ON THE INFINITE SWAPPING LIMIT FOR PARALLEL TEMPERING

PAUL DUPUIS*, YUFEI LIU†, NURIA PLATTNER‡, AND J.D. DOLL§

Abstract. Parallel tempering, also known as replica exchange sampling, is an important method for simulating complex systems. In this algorithm simulations are conducted in parallel at a series of temperatures, and the key feature of the algorithm is a swap mechanism that exchanges configurations between the parallel simulations at a given rate. The mechanism is designed to allow the low temperature system of interest to escape from deep local energy minima where it might otherwise be trapped, via those swaps with the higher temperature components. In this paper we introduce a performance criteria for such schemes based on large deviation theory, and argue that the rate of convergence is a monotone increasing function of the swap rate. This motivates the study of the limit process as the swap rate goes to infinity. We construct a scheme which is equivalent to this limit in a distributional sense, but which involves no swapping at all. Instead, the effect of the swapping is captured by a collection of weights that influence both the dynamics and the empirical measure. While theoretically optimal, this limit is not computationally feasible when the number of temperatures is large, and so variations that are easy to implement and nearly optimal are also developed.

Key words. Markov processes, pure jump, large deviations, relative entropy, ergodic theory, martingale, random measure

AMS subject classifications. 60J25, 60J75, 60F10, 28D20, 60A10, 60G42, 60G57

1. Introduction.

The problem of computing integrals with respect to Gibbs measures occurs in chemistry, physics, statistics, engineering and elsewhere. In many situations, there are no viable alternatives to methods based on Monte Carlo. Given an energy potential, there are standard methods to construct a Markov process whose unique invariant distribution is the associated Gibbs measure, and an approximation is given by the occupation or empirical measure of the process over some finite time interval [10]. However, a weakness of these methods is that they may be slow to converge. This happens when the dynamics of the process do not allow all important parts of the state space to communicate easily with each other. In large scale applications this occurs frequently, since the potential function often has complex structures involving multiple deep local minima.

An interesting method called "parallel tempering" has been designed to overcome some of the difficulties associated with rare transitions [4, 6, 21, 22]. In this technique, simulations are conducted in parallel at a series of temperatures. This method does not require detailed knowledge of or intricate constructions related to the energy surface and is a standard method for simulating complex systems. To illustrate the main idea, we first discuss the diffusion case with two temperatures. Discrete time

*Division of Applied Mathematics, Brown University, Providence, RI 02912. Research supported in part by the Department of Energy (DE-SCOO02413), the National Science Foundation (DMS-1008331), and the Air Force Office of Scientific Research (FA9550-07-1-0544, FA9550-09-1-0378).

†Division of Applied Mathematics, Brown University, Providence, RI 02912. Research supported in part by the Department of Energy (DE-SCOO02413) and the National Science Foundation (DMS-1008331).

‡Department of Chemistry, Brown University, Providence, RI 02912. Research supported in part by the Department of Energy (DE-SCOO02413) and the by postdoctoral support through the Swiss National Science Foundation.

§Department of Chemistry, Brown University, Providence, RI 02912. Research supported in part by the Department of Energy (DE-SCOO02413 and departmental program DE-00015561).

models will be considered later in the paper, and there are obvious analogues for discrete state systems.

Suppose that the probability measure of interest is $\mu(dx) \propto e^{-V(x)/\tau_1}dx$, where τ_1 is the temperature and $V : \mathbb{R}^d \to \mathbb{R}$ is the potential function. The normalization constant of this distribution is typically unknown. Under suitable conditions on V, μ is the unique invariant distribution of the solution to the stochastic differential equation

$$dX = -\nabla V(X)dt + \sqrt{2\tau_1}dW,$$

where W is a d-dimensional standard Wiener process. A straightforward Monte Carlo approximation to μ is the empirical measure over a large time interval of length T, namely,

$$\frac{1}{T}\int_B^{T+B} \delta_{X(t)}(dx)dt,$$

where δ_x is the Dirac measure at x and $B > 0$ denotes a "burn-in" period. When V has multiple deep local minima and the temperature τ_1 is small, the diffusion X can be trapped within these deep local minima for a long time before moving out to other parts of the state space. This is the main cause for the inefficiency.

Now consider a second, larger temperature τ_2. If W_1 and W_2 are independent Wiener processes, then of course the empirical measure of the pair

$$\begin{aligned} dX_1 &= -\nabla V(X_1)dt + \sqrt{2\tau_1}dW_1 \\ dX_2 &= -\nabla V(X_2)dt + \sqrt{2\tau_2}dW_2 \end{aligned} \tag{1.1}$$

gives an approximation to the Gibbs measure with density

$$\pi(x_1, x_2) \propto e^{-\frac{V(x_1)}{\tau_1}} e^{-\frac{V(x_2)}{\tau_2}}. \tag{1.2}$$

The idea of parallel tempering is to allow "swaps" between the components X_1 and X_2. In other words, at random times the X_1 component is moved to the current location of the X_2 component, and vice versa. Swapping is done according to a state dependent intensity, and so the resulting process is actually a Markov jump diffusion. The form of the jump intensity can be selected so that the invariant distribution remains the same, and thus the empirical measure of X_1 can still be used to approximate μ. Specifically, the jump intensity or swapping intensity is of the Metropolis form $ag(X_1, X_2)$, where

$$g(x_1, x_2) = 1 \wedge \frac{\pi(x_2, x_1)}{\pi(x_1, x_2)} \tag{1.3}$$

and $a \in (0, \infty)$ is a constant. Note that the calculation of g does not require the knowledge of the normalization constant. A straightforward calculation shows that π is the stationary density for the resulting process for all values of a [see (2.2)]. We refer to a as the "swap rate," and note that as a increases, the swaps become more frequent.

The intuition behind parallel tempering is that the higher temperature component, being driven by a Wiener process with greater volatility, will move more easily between the different parts of the state space. This "ease-of-movement" is transferred to the lower temperature component via the swapping mechanism so that it is less

likely to be trapped in the deep local minima of the energy potential. This, in turn, is expected to lead to more rapid convergence of the empirical measure to the invariant distribution of the low temperature component. There is an obvious extension to more than two temperatures.

Although this procedure is remarkably simple and needs little detailed information for implementation, relatively little is known regarding theoretical properties. A number of papers discuss the efficiency and optimal design of parallel tempering [8, 16, 15]. However, most of these discussions are based on heuristics and empirical evidence. In general, some care is required to construct schemes that are effective. For example, it can happen that for a given energy potential function and swapping rate, the probability for swapping may be so low that it does not significantly improve performance.

There are two aims to the current paper. The first is to introduce a performance criteria for Monte Carlo schemes of this general kind that differs in some interesting ways from traditional criteria, such as the magnitude of the sub-dominant eigenvalue of a related operator [11, 25]. More precisely, we use the theory of large deviations to define a "rate of convergence" for the empirical measure. The key observation here is that this rate, and hence the performance of parallel tempering, is monotonically increasing with respect to the swap rate a. Traditional wisdom in the application of parallel tempering has been that one should not attempt to swap too frequently. While an obvious reason is that the computational cost for swapping attempts might become a burden, it was also argued that frequent swapping would result in poor sampling. For a discussion on prior approaches to the question of how to set the swapping rate and an argument in favor of frequent swapping, see [20, 19].

The use of this large deviation criteria and the resulting monotonicity with respect to a directly suggest the second aim, which is to study parallel tempering in the limit as $a \to \infty$. Note that the computational cost due just to the swapping will increase without bound, even on bounded time intervals, when $a \to \infty$. However, we will construct an alternative scheme, which uses different process dynamics and a weighted empirical measure. Because this process no longer swaps particle positions, it and the weighted empirical measure have a well-defined limit as $a \to \infty$ which we call infinite swapping. In effect, the swapping is achieved through the proper choice of weights and state dependent diffusion coefficients. This is done for the case of both continuous and discrete time processes with multiple temperatures.

An outline of the paper is as follows. In the next section the swapping model in continuous time is introduced and the rate of convergence, as measured by a large deviations rate function, is defined. The alternative scheme which is equivalent to swapping but which has a well defined limit is introduced, and its limit as $a \to \infty$ is identified. The following section considers the analogous limit model for more than two temperatures, and discusses certain practical difficulties associated with direct implementation when the number of temperatures is not small. The continuous time model is used for illustration because both the large deviation rate and the weak limit of the appropriately redefined swapping model take a simpler form than those of discrete time models. However, the discrete time model is what is actually implemented in practice. To bridge the gap between continuous time diffusion models and discrete time models, in Section 4 we discuss the idea of infinite swapping for continuous time Markov jump processes and prove that the key properties demonstrated for diffusion models hold here as well. We also state a uniform (in the swapping parameter) large deviation principle. The discrete time form actually used in numerical implementa-

tion is presented in Section 5. Section 6 returns to the issue of implementation when the number of temperatures is not small. In particular, we resolve the difficulty of direct implementation of the infinite swapping models via approximation by what we call partial infinite swapping models. Section 7 gives numerical examples, and an appendix gives the proof of the uniform large deviation principle.

2. Diffusion models with two temperatures. Although the implementation of parallel tempering uses a discrete time model, the motivation for the infinite swapping limit is best illustrated in the setting where the state process is a continuous time diffusion process. It is in this case that the large deviation rate function, as well as the construction of a process that is distributionally equivalent to the infinite swapping limit, is simplest. In order to minimize the notational overhead, we discuss in detail the two temperature case. The extension to models with multiple temperatures is obvious and will be stated in the next section.

2.1. Model setup. Let $(\bar{X}_1^a, \bar{X}_2^a)$ denote the Markov jump diffusion process of parallel tempering with swap rate a. That is, between swaps (or jumps), the process follows the diffusion dynamics (1.1). Jumps occur according to the state dependent intensity function $ag(\bar{X}_1^a, \bar{X}_2^a)$. At a jump time t, the particles swap locations, that is, $(\bar{X}_1^a(t), \bar{X}_2^a(t)) = (\bar{X}_2^a(t-), \bar{X}_1^a(t-))$. Hence for a smooth functions $f : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ the infinitesimal generator of the process is given by

$$\begin{aligned}\mathcal{L}^a f(x_1, x_2) = &- \langle \nabla_{x_1} f(x_1, x_2), \nabla V(x_1) \rangle - \langle \nabla_{x_2} f(x_1, x_2), \nabla V(x_2) \rangle \\ &+ \tau_1 \mathrm{tr}\left[\nabla^2_{x_1 x_1} f(x_1, x_2)\right] + \tau_2 \mathrm{tr}\left[\nabla^2_{x_2 x_2} f(x_1, x_2)\right] \\ &+ ag(x_1, x_2)\left[f(x_2, x_1) - f(x_1, x_2)\right],\end{aligned}$$

where $\nabla_{x_i} f$ and $\nabla^2_{x_i x_i} f$ denote the gradient and the Hessian matrix with respect to x_i, respectively, and tr denotes trace. Throughout the paper we also assume the growth condition

$$\lim_{r \to \infty} \inf_{x : |x| \geq r} \langle \nabla V(x), x/|x| \rangle = \infty. \tag{2.1}$$

This condition not only ensures the existence and uniqueness of the invariant distribution, but also enforces the exponential tightness needed for the large deviation principle for the empirical measures.

Recall the definition of π in (1.2) and let μ be the corresponding Gibbs probability distribution, that is,

$$\mu(dx_1 dx_2) = \pi(x_1, x_2) dx_1 dx_2 \propto e^{-\frac{V(x_1)}{\tau_1}} e^{-\frac{V(x_2)}{\tau_2}} dx_1 dx_2.$$

Straightforward calculations show that for any smooth function f which vanishes at infinity

$$\int_{\mathbb{R}^d \times \mathbb{R}^d} \mathcal{L}^a f(x_1, x_2) \mu(dx_1 dx_2) = 0. \tag{2.2}$$

Since the condition (2.1) implies that $V(x) \to \infty$ as $|x| \to \infty$, by the Echeverria's Theorem [5, Theorem 4.9.17], μ is the unique invariant probability distribution of the process $(\bar{X}_1^a, \bar{X}_2^a)$.

2.2. Rate of convergence by large deviations. It follows from the previous discussion and the ergodic theorem [1] that, for a fixed burn-in time B, with probability one

$$\lambda_T^a \doteq \frac{1}{T}\int_B^{T+B} \delta_{(\bar{X}_1^a(t),\bar{X}_2^a(t))}dt \Rightarrow \mu$$

as $T \to \infty$. For notational simplicity we assume without loss of generality that $B = 0$ from now on. A basic question of interest is how rapid is this convergence, and how does it depend on the swap rate a? In particular, what is the rate of convergence of the lower temperature marginal?

We note that standard measures one might use for the rate of convergence, such as the second eigenvalue of the associated operator, are not necessarily appropriate here. They only provide indirect information on the convergence properties of the empirical measure, which is the quantity of interest in the Monte Carlo approximation. Such measures properly characterize the convergence rate of the transition probability

$$p(\boldsymbol{x}, d\boldsymbol{y}; t) = P\left\{(\bar{X}_1^a(t), \bar{X}_2^a(t)) \in d\boldsymbol{y} | (\bar{X}_1^a(0), \bar{X}_2^a(0)) = \boldsymbol{x}\right\}, \quad \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d \times \mathbb{R}^d,$$

as $t \to \infty$. However, they neglect the time averaging effect of the empirical measure, an effect that is not present with the transition probability. In fact, it is easy to construct examples such as nearly periodic Markov chains for which the second eigenvalue suggests a slow convergence when in fact the empirical measure converges quickly [17].

Another commonly used criterion for algorithm performance is the notion of asymptotic variance [10, 12, 23]. For a given functional $f : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, one can establish a central limit theorem which asserts that as $T \to \infty$

$$\mathrm{Var}\left[\frac{1}{\sqrt{T}}\int_0^T f(\bar{X}_1^a(t), \bar{X}_2^a(t))\, dt\right] \to \sigma^2.$$

The magnitude of σ is used to measure the statistical efficiency of the algorithm. The asymptotic variance is closely related to the spectral properties of the underlying probability transition kernel [7, 17]. However, as with the second eigenvalue the usefulness of this criterion for evaluating performance of the empirical measure λ_T^a is not clear.

In this paper, we use the large deviation rate function to characterize the rate of convergence of a sequence of random probability measures. To be more precise, let S be a Polish space, that is, a complete and separable metric space. Denote by $\mathcal{P}(S)$ the space of all probability measures on S. We equip $\mathcal{P}(S)$ with the topology of weak convergence, though one can often use the stronger τ-topology [3]. Under the weak topology, $\mathcal{P}(S)$ is metrizable and itself a Polish space. Note that the empirical measure λ_T^a is a random probability measure, that is, a random variable taking values in the space $\mathcal{P}(S)$.

DEFINITION 2.1. *A sequence of random probability measures $\{\gamma_T\}$ is said to satisfy a large deviation principle (LDP) with rate function $I : \mathcal{P}(S) \to [0, \infty]$, if for all open sets $O \subset \mathcal{P}(S)$*

$$\liminf_{T\to\infty} \frac{1}{T} \log P\{\gamma_T \in O\} \geq -\inf_{\nu \in O} I(\nu),$$

for all closed sets $F \subset \mathcal{P}(S)$

$$\limsup_{T\to\infty} \frac{1}{T} \log P\{\gamma_T \in F\} \le -\inf_{\nu\in F} I(\nu),$$

and if $\{\nu : I(\nu) \le M\}$ *is compact in* $\mathcal{P}(S)$ *for all* $M < \infty$.

For our problem all rate functions encountered will vanish only at the unique invariant distribution μ, and hence give information on the rate of convergence of λ_T^a. A larger rate function will indicate faster convergence, though this is only an asymptotic statement valid for sufficiently large T.

2.3. Explicit form of rate function. The large deviation theory for the empirical measure of a Markov process was first studied in [2]. Besides the Feller property, which will hold for all processes we consider, the validity of the LDP depends on two types of conditions. One is a so-called transitivity condition, which requires that there are times T_1 and T_2 such that for any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d \times \mathbb{R}^d$,

$$\int_0^{T_1} e^{-t} p(\boldsymbol{x}, d\boldsymbol{z}; t)dt \ll \int_0^{T_2} e^{-t} p(\boldsymbol{y}, d\boldsymbol{z}; t)dt,$$

where $\ll$ indicates that the measure in $\boldsymbol{z}$ on the left is absolutely continuous with respect to the measure on the right. For the jump diffusion process we consider here, this condition holds automatically since ∇V is bounded on bounded sets, g is bounded, and the diffusion coefficients are uniformly non-degenerate. The second type of condition is one that enforces a strong form of tightness, such as (2.1).

Under condition (2.1), the LDP holds for $\{\lambda_T^a : T > 0\}$ and the rate function, denoted by I^a, takes a fairly explicit form because the process is in continuous time and reversible [2, 13]. We will state the following result and omit the largely straightforward calculation since its role here is motivational. [A uniform LDP for the analogous jump Markov process will be stated in Section 4, and its proof is given in the appendix.]

Let ν be a probability measure on $\mathbb{R}^d \times \mathbb{R}^d$ with smooth density. Define $\theta(x_1, x_2) \doteq [d\nu/d\mu](x_1, x_2)$. Then $I^a(\nu)$ can be expressed as

$$I^a(\nu) = J_0(\nu) + aJ_1(\nu), \tag{2.3}$$

where

$$J_0(\nu) = \int_{\mathbb{R}^d\times\mathbb{R}^d} \frac{1}{8\theta(x_1,x_2)^2} \left[\tau_1 \left\|\nabla_{x_1}\theta(x_1,x_2)\right\|^2 + \tau_2 \left\|\nabla_{x_2}\theta(x_1,x_2)\right\|^2\right] \nu(dx_1dx_2)$$

$$J_1(\nu) = \int_{\mathbb{R}^d\times\mathbb{R}^d} g(x_1,x_2)\ell\left(\sqrt{\frac{\theta(x_2,x_1)}{\theta(x_1,x_2)}}\right) \nu(dx_1dx_2),$$

and where $\ell(z) = z\log z - z + 1$ for $z \ge 0$ is familiar from the large deviation theory for jump processes.

The key observation is that the rate function $I^a(\nu)$ is affine in the swapping rate a, with $J_0(\nu)$ the rate function in the case of no swapping. Furthermore, $J_1(\nu) \ge 0$ with equality if and only if $\theta(x_2, x_1) = \theta(x_1, x_2)$ for ν-a.e. (x_1, x_2). This form of the rate function, and in particular its monotonicity in a, motivates the study of the *infinite swapping limit* as $a \to \infty$.

REMARK 2.2. The limit of the rate function I^a satisfies

$$I^\infty(\nu) \doteq \lim_{a\to\infty} I^a(\nu) = \begin{cases} J_0(\nu) & \theta(x_1,x_2) = \theta(x_2,x_1)\ \nu\text{-a.s.}, \\ \infty & \textit{otherwise.} \end{cases}$$

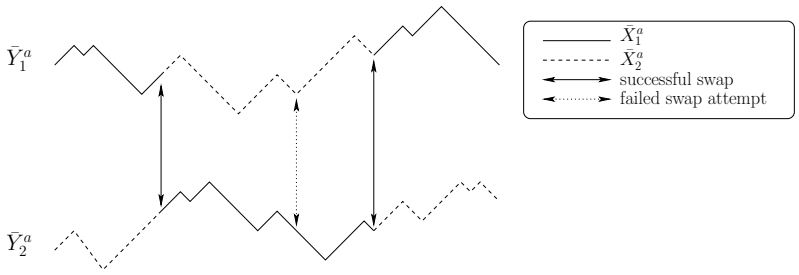

FIG. 2.1. *Temperature swapped and particle swapped processes*

Hence for $I^\infty(\nu)$ to be finite it is necessary that ν put exactly the same relative weight as μ on the points (x_1, x_2) and (x_2, x_1). Note that if a process could be constructed with I^∞ as its rate function, then with the large deviation rate as our criteria such a process improves on parallel tempering with finite swapping rate in exactly those situations where parallel tempering improves on the process with no swapping at all.

2.4. Infinite swapping limit. From a practical perspective, it may appear that there are limitations on how much benefit one obtains by letting $a \to \infty$. When implemented in discrete time, the overall jump intensity corresponds to the generation of roughly a independent random variables that are uniform on $[0, 1]$ for each corresponding unit of continuous time, and based on each uniform variable a comparison is made to decide whether or not to make the swap. Hence even for fixed and finite T, the computations required to simulate a single trajectory scale like a as $a \to \infty$. Thus it is of interest if one can gain the benefit of the higher swapping rate without all the computational burden. This turns out to be possible, but requires that we view the prelimit processes in a different way.

It is clear that the processes $(\bar{X}_1^a, \bar{X}_2^a)$ are not tight as $a \to \infty$, since the number of discontinuities of size $O(1)$ will grow without bound in any time interval of positive length. In order to obtain a limit, we consider alternative processes defined by

$$\begin{aligned} d\bar{Y}_1^a &= -\nabla V(\bar{Y}_1^a)dt + \sqrt{2\tau_1 1_{\{\bar{Z}^a=0\}} + 2\tau_2 1_{\{\bar{Z}^a=1\}}}\, dW_1 \\ d\bar{Y}_2^a &= -\nabla V(\bar{Y}_2^a)dt + \sqrt{2\tau_2 1_{\{\bar{Z}^a=0\}} + 2\tau_1 1_{\{\bar{Z}^a=1\}}}\, dW_2 \end{aligned} \tag{2.4}$$

where $\bar{Z}^a$ is a jump process that switches from state 0 to state 1 with intensity $ag(\bar{Y}_1^a, \bar{Y}_2^a)$ and from state 1 to state 0 with intensity $ag(\bar{Y}_2^a, \bar{Y}_1^a)$. Compared to conventional parallel tempering, the processes $(\bar{Y}_1^a, \bar{Y}_2^a)$ swap the diffusion coefficients at the jump times rather than the physical locations of two particles with constant diffusion coefficients. For this reason, we refer to the solution to (2.4) as the *temperature swapped process,* in order to distinguish it from the *particle swapped process* $(\bar{X}_1^a, \bar{X}_2^a)$. We illustrate these processes in Figure 2.1. Note that the solid line and the dotted line represent $\bar{X}_1^a$ and $\bar{X}_2^a$, respectively. These processes have more and more frequent jumps of size $O(1)$ as $a \to \infty$. In contrast, the process $(\bar{Y}_1^a, \bar{Y}_2^a)$ have varying diffusion coefficient. The figure attempts to also suggest features of the discrete time setting, with both successful and failed swap attempts.

Clearly the empirical measure of $(\bar{Y}_1^a, \bar{Y}_2^a)$ does not provide an approximation to μ. Instead, we should shift attention between $(\bar{Y}_1^a, \bar{Y}_2^a)$ and $(\bar{Y}_2^a, \bar{Y}_1^a)$ depending on

the value of $\bar{Z}^a$. Indeed, the random probability measures

$$\eta_T^a = \frac{1}{T}\int_0^T \left[1_{\{\bar{Z}^a(t)=0\}}\delta_{(\bar{Y}_1^a(t),\bar{Y}_2^a(t))} + 1_{\{\bar{Z}^a(t)=1\}}\delta_{(\bar{Y}_2^a(t),\bar{Y}_1^a(t))}\right] dt \tag{2.5}$$

have the same distribution as

$$\frac{1}{T}\int_0^T \delta_{(\bar{X}_1^a(s),\bar{X}_2^a(s))} ds,$$

and hence converge to μ at the same rate. However, these processes and measures have well defined limits in distribution as $a \to \infty$. More precisely, we have the following result. For the proof see [9]. Related (but more complex) calculations are needed to prove the uniform large deviation result given in Theorem 4.1.

THEOREM 2.3. *Assume that ∇V is locally Lipschitz continuous. Then for each T the sequence $(\bar{Y}_1^a, \bar{Y}_2^a, \eta_T^a)$ converges in distribution to $(\bar{Y}_1^\infty, \bar{Y}_2^\infty, \eta_T)$ as $a \to \infty$, where $(\bar{Y}_1^\infty, \bar{Y}_2^\infty)$ is the unique strong solution to*

$$\begin{aligned} d\bar{Y}_1^\infty &= -\nabla V(\bar{Y}_1^\infty)dt + \sqrt{2\tau_1\rho(\bar{Y}_1^\infty, \bar{Y}_2^\infty) + 2\tau_2\rho(\bar{Y}_2^\infty, \bar{Y}_1^\infty)}\, dW_1 \\ d\bar{Y}_2^\infty &= -\nabla V(\bar{Y}_2^\infty)dt + \sqrt{2\tau_2\rho(\bar{Y}_1^\infty, \bar{Y}_2^\infty) + 2\tau_1\rho(\bar{Y}_2^\infty, \bar{Y}_1^\infty)}\, dW_2, \end{aligned} \tag{2.6}$$

$$\eta_T^\infty = \frac{1}{T}\int_0^T \left[\rho(\bar{Y}_1^\infty(t), \bar{Y}_2^\infty(t))\delta_{(\bar{Y}_1^\infty(t),\bar{Y}_2^\infty(t))} + \rho(\bar{Y}_2^\infty(t), \bar{Y}_1^\infty(t))\delta_{(\bar{Y}_2^\infty(t),\bar{Y}_1^\infty(t))}\right] dt, \tag{2.7}$$

and

$$\rho(x_1, x_2) \doteq \frac{\pi(x_1, x_2)}{\pi(x_2, x_1) + \pi(x_1, x_2)}.$$

The existence and form of the limit are due to the time scale separation between the fast $\bar{Z}^a$ process and the slow $(\bar{Y}_1^a, \bar{Y}_2^a)$ process. To give an intuitive explanation of the limit dynamics, consider the prelimit processes (2.4). Suppose that on a small time interval, the value of the slow process $(\bar{Y}_1^a, \bar{Y}_2^a)$ does not vary much, say $(\bar{Y}_1^a, \bar{Y}_2^a) \approx (x_1, x_2)$. Given the dynamics of the binary process $\bar{Z}^a$, it is easy to verify that as a tends to infinity the fractions of time that $\bar{Z}^a = 0$ and $\bar{Z}^a = 1$ are $\rho(x_1, x_2)$ and $\rho(x_2, x_1)$, respectively. This leads to the limit dynamics (2.6). When mapped back to the particle swapped process, $\rho(x_1, x_2)$ and $\rho(x_2, x_1)$ account for the fraction of time that $(\bar{X}_1^a, \bar{X}_2^a) = (x_1, x_2)$ and $(\bar{X}_1^a, \bar{X}_2^a) = (x_2, x_1)$, respectively, which naturally leads to the limit weighted empirical measure (2.7).

The weights ρ_1 and ρ_2 do not depend on the unknown normalization constant, and in fact

$$\rho(x_1, x_2) = \frac{e^{-\frac{V(x_1)}{\tau_1} - \frac{V(x_2)}{\tau_2}}}{e^{-\frac{V(x_1)}{\tau_1} - \frac{V(x_2)}{\tau_2}} + e^{-\frac{V(x_2)}{\tau_1} - \frac{V(x_1)}{\tau_2}}} \tag{2.8}$$

and

$$\rho(x_2, x_1) = 1 - \rho(x_1, x_2) = \frac{e^{-\frac{V(x_2)}{\tau_1} - \frac{V(x_1)}{\tau_2}}}{e^{-\frac{V(x_1)}{\tau_1} - \frac{V(x_2)}{\tau_2}} + e^{-\frac{V(x_2)}{\tau_1} - \frac{V(x_1)}{\tau_2}}}.$$

The following properties of the limit system are worth noting.

1. Instantaneous equilibration of multiple locations. Observe that the lower temperature component of this modified "empirical measure," i.e., the first marginal, uses contributions from both components at all times, corrected according to the weights. The form of the weights in (2.7) guarantees that the contributions to η_T^∞ from locations $(\bar{Y}_1^\infty, \bar{Y}_2^\infty)$ and $(\bar{Y}_2^\infty, \bar{Y}_1^\infty)$ are at any time perfectly balanced according to the invariant distribution on product space.
2. Symmetry and invariant distribution. While the marginals of η_T^∞ play very different roles, the dynamics of $\bar{Y}_1^\infty$ and $\bar{Y}_2^\infty$ are actually symmetric. Using the Echeverria's Theorem [5, Theorem 4.9.17], it can be shown that the unique invariant distribution of the process $(\bar{Y}_1^\infty, \bar{Y}_2^\infty)$ has the density

$$\frac{1}{2}[\pi(x_1, x_2) + \pi(x_2, x_1)].$$

   It then follows from the ergodic theorem that $\eta_T^\infty \Rightarrow \mu$ w.p.1 as $T \to \infty$. This is hardly surprising, since μ is the invariant distribution for the prelimit processes $(\bar{X}_1^a, \bar{X}_2^a)$.
3. Escape from local minima. Finally it is worth commenting on the behavior of the diffusion coefficients as a function of the relative positions of $\bar{Y}_1^\infty$ and $\bar{Y}_2^\infty$ on an energy landscape. Recall that $\tau_1 < \tau_2$. Suppose that $\bar{Y}_1^\infty(t)$ is near the bottom of a local minimum (which for simplicity we set to be zero), while $\bar{Y}_2^\infty(t)$ is at a higher energy level, perhaps within the same local minimum. Then

$$\rho(y_1, y_2) \approx \frac{e^{-\frac{V(y_2)}{\tau_2}}}{e^{-\frac{V(y_2)}{\tau_2}} + e^{-\frac{V(y_2)}{\tau_1}}} \approx 1, \qquad \rho(y_2, y_1) = 1 - \rho(y_1, y_2) \approx 0.$$

   Thus to some degree the dynamics look like

$$d\bar{Y}_1^\infty = -\nabla V(\bar{Y}_1^\infty)dt + \sqrt{2\tau_1}dW_1$$
$$d\bar{Y}_2^\infty = -\nabla V(\bar{Y}_2^\infty)dt + \sqrt{2\tau_2}dW_2,$$

   i.e., the particle higher up on the energy landscape is given the greater diffusion coefficient, while the one near the bottom of the well is automatically given the lower coefficient. Hence the particle which is already closer to escaping from the well is automatically given the greater noise (within the confines of (τ_1, τ_2)). Recalling the role of the higher temperature particle is to more assiduously explore the landscape in parallel tempering, this is an interesting property.

One can apply results from [2] to show that the empirical measure of the infinite swapping limit $\{\eta_T^\infty : T > 0\}$ satisfies a large deviation principle with rate function I^∞ as defined in Remark 2.2. However, to justify the claim that the infinite swapping model is truly superior to the finite swapping variant (note that $I^\infty \geq I^a$ for any finite a), one should establish a *uniform* large deviation principle, which would show that I^∞ is the correct rate function for any sequence $\{a_T : T > 0\} \subset [0, \infty]$ such that $a_T \to \infty$ as $T \to \infty$. We omit the proof here, since in Theorem 4.1 the analogous result will be proved in the setting of continuous time jump Markov processes.

3. Diffusion models with multiple temperatures. In practice parallel tempering uses swaps between more than two temperatures. A key reason is that if the

gap between the temperatures is too large then the probability of a successful swap under the [discrete time version of the] Metropolis rule (1.3) is far too low for the exchange of information to be effective. A natural generalization is to introduce, to the degree that computational feasibility is maintained, a ladder of higher temperatures, and then attempt pairwise swaps between particles. There are a variety of schemes used to select which pair to attempt the swap, including deterministic and randomized rules for selecting only adjacent temperatures or arbitrary pair of temperatures. However, if one were to replace any of these particle swapped processes with its equivalent temperature swapped analogue and consider the infinite swapping limit, one would get the same system of process dynamics and weighted empirical measures which we now describe.

Suppose that besides the lowest temperature τ_1 (in many cases the temperature of principal interest), we introduce the collection of higher temperatures

$$\tau_1 < \tau_2 < \cdots < \tau_K.$$

Let $\boldsymbol{y} = (y_1, y_2, \ldots, y_K) \in (\mathbb{R}^d)^K$ be a generic point in the state space of the process and define a product Gibbs distribution with the density

$$\pi(\boldsymbol{y}) = \pi(y_1, y_2, \ldots, y_K) \propto e^{-V(y_1)/\tau_1} e^{-V(y_2)/\tau_2} \cdots e^{-V(y_K)/\tau_K}.$$

The limit of the temperature swapped processes with K temperatures takes the form

$$\begin{aligned}
d\bar{Y}_1^\infty &= -\nabla V\left(\bar{Y}_1^\infty\right) dt + \sqrt{2\rho_{11}\tau_1 + 2\rho_{12}\tau_2 + \ldots + 2\rho_{1K}\tau_K}\, dW_1 \\
d\bar{Y}_2^\infty &= -\nabla V\left(\bar{Y}_2^\infty\right) dt + \sqrt{2\rho_{21}\tau_1 + 2\rho_{22}\tau_2 + \ldots + 2\rho_{2K}\tau_K}\, dW_2 \\
&\vdots \\
d\bar{Y}_K^\infty &= -\nabla V\left(\bar{Y}_K^\infty\right) dt + \sqrt{2\rho_{K1}\tau_1 + 2\rho_{K2}\tau_2 + \ldots + 2\rho_{KK}\tau_K}\, dW_K.
\end{aligned}$$

To define these weights ρ_{ij} it is convenient to introduce some new notation.

Let S_K be the collection of all bijective mappings from $\{1, 2, \ldots, K\}$ to itself. S_K has $K!$ elements, each of which corresponds to a unique permutation of the set $\{1, 2, \ldots, K\}$, and S_K forms a group with the group action defined by composition. Let σ^{-1} denote the inverse of σ. Furthermore, for each $\sigma \in S_K$ and every $\boldsymbol{y} = (y_1, y_2, \ldots, y_K) \in (\mathbb{R}^d)^K$, define $\boldsymbol{y}_\sigma \doteq \left(y_{\sigma(1)}, y_{\sigma(2)}, \ldots, y_{\sigma(K)}\right)$.

At the level of the prelimit particle swapped process, we interpret the permutation σ to correspond to event that the particles at location $\boldsymbol{y} = (y_1, y_2, \ldots, y_K)$ are swapped to the new location $\boldsymbol{y}_\sigma = \left(y_{\sigma(1)}, y_{\sigma(2)}, \ldots, y_{\sigma(K)}\right)$. Under the temperature swapped process, this corresponds to the event that particles initially assigned temperatures in the order $\tau_1, \tau_2, \ldots, \tau_K$ have now been assigned the temperatures $\tau_{\sigma^{-1}(1)}, \tau_{\sigma^{-1}(2)}, \ldots, \tau_{\sigma^{-1}(K)}$.

The identification of the infinite swapping limit of the temperature swapped processes is very similar to that of the two temperature model in the previous section. By exploiting the time-scale separation, one can assume that in a small time interval the only motion is due to temperature swapping and the motion due to diffusion is negligible. Hence the fraction of time that the permutation σ is in effect should again be proportional to the relative weight assigned by the invariant distribution to $\boldsymbol{y}_\sigma$, that is,

$$\pi(\boldsymbol{y}_\sigma) = \pi\left(y_{\sigma(1)}, y_{\sigma(2)}, \ldots, y_{\sigma(K)}\right).$$

Thus if

$$w(\boldsymbol{y}) \doteq \frac{\pi(\boldsymbol{y})}{\sum_{\theta \in S_K} \pi(\boldsymbol{y}_\theta)},$$

then the fraction of time that the permutation σ is in effect is $w(\boldsymbol{y}_\sigma)$. Note that for any $\boldsymbol{y}$,

$$\sum_{\sigma \in S_K} w(\boldsymbol{y}_\sigma) = 1.$$

Going back to the definition of the weights $\rho_{ij}(\boldsymbol{y})$, $i, j = 1, \ldots, K$, it is clear that they represent the limit proportion of time that the i-th particle is assigned the temperature j and hence will satisfy

$$\rho_{ij}(\boldsymbol{y}) = \sum_{\sigma:\ \sigma(j)=i} w(\boldsymbol{y}_\sigma).$$

Likewise the replacement for the empirical measure, accounting as it does for mapping the temperature swapped process back to the particle swapped process, is given by

$$\eta_T^\infty = \frac{1}{T} \int_0^T \sum_{\sigma \in S_K} w(\bar{\boldsymbol{Y}}_\sigma^\infty(t)) \delta_{\bar{\boldsymbol{Y}}_\sigma^\infty(t)} dt, \tag{3.1}$$

where $\bar{\boldsymbol{Y}}_\sigma^\infty(t) \doteq [\bar{\boldsymbol{Y}}^\infty(t)]_\sigma = (\bar{Y}_{\sigma(1)}^\infty(t), \bar{Y}_{\sigma(2)}^\infty(t), \ldots, \bar{Y}_{\sigma(K)}^\infty(t))$.

The instantaneous equilibration property still holds for the infinite swapping system with multiple temperatures. That is, at any time $t \in [0, T]$ and given a current position $\bar{\boldsymbol{Y}}^\infty(t) = \boldsymbol{y}$, the weighted empirical measure η_T^∞ has contributions from all locations of the form $\boldsymbol{y}_\sigma, \sigma \in S_K$, balanced exactly according to their relative contributions from the invariant density $\pi(\boldsymbol{y}_\sigma)$. The dynamics of $\bar{\boldsymbol{Y}}^\infty$ are again symmetric, and the density of the invariant distribution at point $\boldsymbol{y}$ is

$$\frac{1}{K!} \sum_{\sigma \in S_K} \pi(\boldsymbol{y}_\sigma).$$

REMARK 3.1. The infinite swapping process described above allows the most effective communication between all temperatures, and is the "best" in the sense that it leads to the largest large deviation rate function and hence the fastest rate of convergence. However, computation of the coefficients becomes very demanding for even moderate values of K, since one needs to evaluate $K!$ terms from all possible permutations. In Section 5 we discuss a more tractable and easily implementable family of schemes which are essentially approximations to the infinite swapping model presented in the current section and have very similar performance. We call the current model the full infinite swapping model since it uses the whole permutation group S_K, as opposed to the partial infinite swapping model in Section 5 where only subgroups of S_K are used.

4. Infinite swapping for jump Markov processes. The continuous time diffusion model is a convenient vehicle to convey the main idea of infinite swapping. In practice, however, algorithms are implemented in discrete time. In this section we discuss continuous time pure jump Markov processes and the associated infinite swapping limit. The purpose of this intermediate step is to serve as a bridge between the

diffusion and discrete time Markov chain models. These two types of processes have some subtle differences regarding the infinite swapping limit which is best illustrated through the continuous time jump Markov model.

In this section we discuss the two-temperature model, and omit the completely analogous multiple-temperature counterpart. We will not refer to temperatures τ_1 and τ_2 to distinguish dynamics. Instead, let $\alpha_1(x, dy)$ and $\alpha_2(x, dy)$ be two probability transition kernels on $\mathbb{R}^d$ given $\mathbb{R}^d$. One can think of α_i as the dynamics under temperature τ_i for $i = 1, 2$. We assume that for each $i = 1, 2$ the stationary distribution μ_i associated with the transition kernel α_i admits the density π_i in order to be consistent with the diffusion models, and define

$$\mu = \mu_1 \times \mu_2, \qquad \pi(x_1, x_2) = \pi_1(x_1)\pi_2(x_2).$$

We assume that the kernels are Feller and have a density that is uniformly bounded with respect to Lebesgue measure. These conditions would always be satisfied in practice. Finally, we assume that the detailed balance or reversibility condition holds, that is,

$$\alpha_i(x, dz)\pi_i(x)dx = \alpha_i(z, dx)\pi_i(z)dz. \tag{4.1}$$

4.1. Model setup. In the absence of swapping [i.e., swapping rate $a = 0$], the dynamics of the system are as follows. Let $\boldsymbol{X}^0 = \{\boldsymbol{X}^0(t) = (X_1^0(t), X_2^0(t)) : t \geq 0\}$ denote a continuous time process taking values in $\mathbb{R}^d \times \mathbb{R}^d$. The probability transition kernel associated with the embedded Markov chain, denoted by $\bar{\boldsymbol{X}}^0 = \{(\bar{X}_1^0(j), \bar{X}_2^0(j)) : j = 0, 1, \ldots\}$, is

$$P\{\bar{\boldsymbol{X}}^0(j+1) \in (dy_1, dy_2) | \bar{\boldsymbol{X}}^0(j) = (x_1, x_2)\} = \alpha_1(x_1, dy_1)\alpha_2(x_2, dy_2).$$

Without loss of generality, we assume that the jump times occur according to a Poisson process with rate one. In other words, let $\{\tau_i\}$ be a sequence of independent and identically distributed (iid) exponential random variables with rate one that are independent of $\bar{\boldsymbol{X}}^0$. Then

$$\boldsymbol{X}^0(t) = \bar{\boldsymbol{X}}^0(j), \quad \text{for } \sum_{i=1}^{j} \tau_i \leq t < \sum_{i=1}^{j+1} \tau_i.$$

The infinitesimal generator of $\boldsymbol{X}^0$ is such that for a given smooth function f,

$$\mathcal{L}^0 f(x_1, x_2) = \int_{\mathbb{R}^d \times \mathbb{R}^d} [f(y_1, y_2) - f(x_1, x_2)]\alpha_1(x_1, dy_1)\alpha_2(x_2, dy_2).$$

Owing to the detailed balance condition (4.1), the operator $\mathcal{L}^0$ is self-adjoint. Using arguments similar to but simpler than those used to prove the uniform LDP in Theorem 4.1, the large deviations rate function I^0 associated with the occupation measure

$$\eta_T^0 = \frac{1}{T} \int_0^T \delta_{\boldsymbol{X}^0(t)}\, dt$$

can be explicitly identified: for any probability measure ν on $\mathbb{R}^d \times \mathbb{R}^d$ with $\nu \ll \mu$ and $\theta = d\nu/d\mu$,

$$I^0(\nu) = 1 - \int_{(\mathbb{R}^d \times \mathbb{R}^d)^2} \sqrt{\theta(x_1, x_2)\theta(y_1, y_2)}\pi(x_1, x_2)\alpha_1(x_1, dy_1)\alpha_2(x_2, dy_2)\, dx_1 dx_2,$$

and I^0 is extended to all of $\mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d)$ by lower semicontinuous regularization.

4.2. Finite swapping model. Denote by $\boldsymbol{X}^a = \{(X_1^a(t), X_2^a(t)) : t \geq 0\}$ the state process of the finite swapping model with swapping rate a, and let $\bar{\boldsymbol{X}}^a = \{(\bar{X}_1^a(j), \bar{X}_2^a(j)) : j = 0, 1, \ldots\}$ be the embedded Markov chain. The probability transition kernel for $\bar{\boldsymbol{X}}^a$ is

$$P\{\bar{\boldsymbol{X}}^a(j+1) \in (dy_1, dy_2) | \bar{\boldsymbol{X}}^a(j) = (x_1, x_2)\} = \frac{1}{a+1}\alpha_1(x_1, dy_1)\alpha_2(x_2, dy_2) + \frac{a}{a+1}\left[g(x_1, x_2)\delta_{(x_2,x_1)}(dy_1, dy_2) + (1 - g(x_1, x_2))\delta_{(x_1,x_2)}(dy_1, dy_2)\right],$$

where g is defined as in (1.3). Furthermore, let $\{\tau_i^a\}$ be a sequence of iid exponential random variables with rate $(a+1)$ and define

$$\boldsymbol{X}^a(t) = \bar{\boldsymbol{X}}^a(j), \quad \text{for } \sum_{i=1}^{j} \tau_i^a \leq t < \sum_{i=1}^{j+1} \tau_i^a.$$

In other words, the jumps occur according to a Poisson process with rate $a+1$. Note that there are two types of jumps. At any jump time, with probability $1/(a+1)$ it will be a jump according to the underlying probability transition kernels α_1 and α_2. With probability $a/(a+1)$ it will be an attempted swap which will succeed with the probability determined by g. As a grows, the swap attempts become more and more frequent. However, the time between two consecutive jumps of the first type will have the same distribution as

$$S^a = \sum_{i=1}^{N^a} \tau_i^a$$

where N^a is a geometric random variable with parameter $1/(a+1)$. It is easy to argue that for any a the distribution of S^a is exponential with rate one. This observation will be useful when we derive the infinite swapping limit.

The infinitesimal generator $\mathcal{L}^a$ of $\bar{\boldsymbol{X}}^a$ is such that for any smooth function f on $\mathbb{R}^d \times \mathbb{R}^d$

$$\mathcal{L}^a f(x_1, x_2) = \int_{\mathbb{R}^d \times \mathbb{R}^d} [f(y_1, y_2) - f(x_1, x_2)]\alpha_1(x_1, dy_1)\alpha_2(x_2, dy_2) + ag(x_1, x_2)[f(x_2, x_1) - f(x_1, x_2)].$$

It is not difficult to check that the stationary distribution of $\bar{\boldsymbol{X}}^a$ remains μ and that $\mathcal{L}^a$ is self-adjoint. As before, the large deviation rate function I^a for the occupation measure

$$\eta_T^a = \frac{1}{T}\int_0^T \delta_{\boldsymbol{X}^a(t)}\, dt \tag{4.2}$$

can be explicitly identified. Indeed, for any probability measure ν on $\mathbb{R}^d \times \mathbb{R}^d$ with $\nu \ll \mu$ and $\theta = d\nu/d\mu$

$$I^a(\nu) = I^0(\nu) + aJ(\nu),$$

where

$$J(\nu) = \int_{\mathbb{R}^d \times \mathbb{R}^d} g(x_1, x_2)\ell\left(\sqrt{\frac{\theta(x_2, x_1)}{\theta(x_1, x_2)}}\right)\nu(dx_1, dx_2).$$

Note that as before, I^a is monotonically increasing with respect to a. Since $J(\nu) \geq 0$ with equality if and only if $\theta(x_1, x_2) = \theta(x_2, x_1)$ ν-a.e., we have

$$I^\infty(\nu) \doteq \lim_{a\to\infty} I^a(\nu) = \begin{cases} I^0(\nu) & \text{if } \theta(x_1, x_2) = \theta(x_2, x_1) \ \nu\text{-a.s.}, \\ \infty & \text{otherwise.} \end{cases} \tag{4.3}$$

4.3. Infinite swapping limit. The infinite swapping limit for $\boldsymbol{X}^a$ as $a \to \infty$ can be similarly obtained by considering the corresponding temperature swapped processes. Since the times between jumps determined by α_1 and α_2 are always exponential with rate one, the infinite swapping limit $\boldsymbol{Y}^\infty = (Y_1^\infty, Y_2^\infty)$ is a pure jump Markov process where jumps occur according to a Poisson process with rate one. In other words,

$$\boldsymbol{Y}^\infty(t) = \bar{\boldsymbol{Y}}^\infty(j), \quad \text{for } \sum_{i=1}^{j} \tau_i \leq t < \sum_{i=1}^{j+1} \tau_i,$$

where $\bar{\boldsymbol{Y}}^\infty$ is the embedded Markov chain and $\{\tau_i\}$ a sequence of iid exponential random variables with rate one. Furthermore, the probability transition kernel for $\bar{\boldsymbol{Y}}^\infty$ is

$$\begin{aligned} &P\{\bar{\boldsymbol{Y}}^\infty(j+1) \in (dz_1, dz_2) | \bar{\boldsymbol{Y}}^\infty(j) = (y_1, y_2)\} \\ &= \rho(y_1, y_2)\alpha_1(y_1, dz_1)\alpha_2(y_2, dz_2) + \rho(y_2, y_1)\alpha_2(y_1, dz_1)\alpha_1(y_2, dz_2), \end{aligned} \tag{4.4}$$

where the weight function ρ is defined as in Theorem 2.3. It is not difficult to argue that the stationary distribution for $\boldsymbol{Y}^\infty$ is

$$\bar{\mu}(dy_1, dy_2) = \frac{1}{2}[\pi(y_1, y_2) + \pi(y_2, y_1)]\, dy_1 dy_2,$$

and the weighted occupation measure

$$\eta_T^\infty = \frac{1}{T}\int_0^T \left[\rho(Y_1^\infty(t), Y_2^\infty(t))\delta_{(Y_1^\infty(t), Y_2^\infty(t))} + \rho(Y_2^\infty(t), Y_1^\infty(t))\delta_{(Y_2^\infty(t), Y_1^\infty(t))}\right] dt \tag{4.5}$$

converges to $\mu(dx_1, dx_2) = \pi(x_1, x_2)dx_1 dx_2$ as $T \to \infty$. It is obvious that the dynamics of the infinite swapping limit are symmetric and instantaneously equilibrate the contribution from (Y_1, Y_2) and (Y_2, Y_1) according to the invariant measure, owing to the weight function ρ.

We have the following uniform large deviation principle result, which justifies the superiority of infinite swapping model. Its proof is deferred to the appendix. It should be noted that rate identification is not covered by the existing literature, even in the case of a fixed swapping rate, due to the pure jump nature of the process.

THEOREM 4.1. *The occupation measure $\{\eta_T^\infty : T > 0\}$ satisfies a large deviation principle with rate function I^∞. More generally, define the finite swapping model as in Subsection 4.2. Consider any sequence $\{a_T : T > 0\} \subset [0, \infty]$ such that $a_T \to \infty$ as $T \to \infty$, and interpret $a_T < \infty$ to mean that $\eta_T^{a_T}$ is defined by (4.2) with $a = a_T$, and $a_T = \infty$ to mean that $\eta_T^{a_T}$ is defined by (4.5). Then $\{\eta_T^{a_T} : T > 0\}$ satisfies an LDP with the rate function I^∞ defined in equation (4.3).*

5. Discrete time process models.

5.1. Conventional parallel tempering algorithms. In the discrete time, multi-temperature algorithms that are actually implemented, a swap is attempted after a deterministic or random number of time steps, with a success probability of the form (1.3). The two temperatures corresponding to particles for which a swap is attempted can be chosen according to a deterministic or random schedule, and as noted previously are usually adjacent since otherwise the success probability (1.3) will be too small to allow efficient exchange of information.

As before it suffices to describe the algorithm in the setting of two temperatures. As in Section 4, let $\alpha_i(x, dy)$ denote the probability transition kernel for temperature τ_i whose stationary distribution has a density π_i for $i = 1, 2$. For now let $N = 1/a$ be a fixed positive integer that determines the frequency of swap attempts. Let $\bar{\boldsymbol{X}} = \{(\bar{X}_1(j), \bar{X}_2(j)) : j = 0, 1, \ldots\}$ denote the state process. Then the evolution of the dynamics is as follows. For any integer $k \geq 1$ and $(k-1)(N+1) \leq j \leq k(N+1)-2$,

$$P\{\bar{\boldsymbol{X}}(j+1) \in (dy_1, dy_2) | \bar{\boldsymbol{X}}(j) = (x_1, x_2)\} = \alpha_1(x_1, dy_1)\alpha_2(x_2, dy_2)$$

and for $j = k(N+1) - 1$,

$$\begin{aligned} P\{\bar{\boldsymbol{X}}(j+1) = (x_2, x_1) | \bar{\boldsymbol{X}}(j) = (x_1, x_2)\} &= g(x_1, x_2), \\ P\{\bar{\boldsymbol{X}}(j+1) = (x_1, x_2) | \bar{\boldsymbol{X}}(j) = (x_1, x_2)\} &= 1 - g(x_1, x_2). \end{aligned}$$

Thus a swap is attempted after every N ordinary time steps based on the underlying transition kernels α_1 and α_2. The case $N = 1/a$ with a an integer greater than one corresponds to the case where multiple swaps are attempted between two ordinary time steps. The unique invariant distribution of $\bar{X}$ is $\mu(dx_1 dx_2) = \pi(x_1, x_2)\, dx_1 dx_2$, regardless of the value of N, and the occupation measure

$$\frac{1}{J} \sum_{j=0}^{J-1} \delta_{\boldsymbol{X}(j)}$$

converges to μ as $J \to \infty$ almost surely.

REMARK 5.1. Note that N could be random. For example, if N is chosen to be a geometric random variable with mean λ, then $\bar{X}$ is exactly the embedded Markov chain of the pure jump Markov process $\bar{X}^a$ with $a = 1/\lambda$ in Subsection 4.2.

5.2. Infinite swapping model. As with the continuous time case, to produce a well-defined limit one must consider the temperature swapped process and then consider the limit as swapping frequency tends to infinity. It turns out that the limit is exactly the embedded Markov chain for the pure jump Markov process in Subsection 4.3. That is, the infinite swapping limit in discrete time is a Markov chain $\bar{\boldsymbol{Y}}^\infty = \left\{(\bar{Y}_1^\infty(j), \bar{Y}_2^\infty(j)) : j = 0, 1, \ldots\right\}$ with the transition kernel

$$\rho(y_1, y_2)\alpha_1(y_1, dz_1)\alpha_2(y_2, dz_2) + \rho(y_2, y_1)\alpha_2(y_1, dz_1)\alpha_1(y_2, dz_2). \tag{5.1}$$

The corresponding weighted empirical measure is

$$\eta_J^\infty \doteq \frac{1}{J} \sum_{j=0}^{J-1} \left[\rho(\bar{Y}_1^\infty(j), \bar{Y}_2^\infty(j))\delta_{(\bar{Y}_1^\infty(j), \bar{Y}_2^\infty(j))} + \rho(\bar{Y}_2^\infty(j), \bar{Y}_1^\infty(j))\delta_{(\bar{Y}_2^\infty(j), \bar{Y}_1^\infty(j))}\right]. \tag{5.2}$$

The generalization to multiple temperatures is also straightforward. Suppose that there are K temperatures. Denote the infinite swapping limit process by $\bar{\boldsymbol{Y}} = \{\bar{\boldsymbol{Y}}(j) :$

$j = 0, 1, \ldots\}$, which is a Markov chain taking values in the space $(\mathbb{R}^d)^K$. Given that the current state of the chain is $\bar{\boldsymbol{Y}}^\infty(j) = \boldsymbol{y} = (y_1, \ldots, y_K)$, define as before the weights

$$w(\boldsymbol{y}) \doteq \frac{\pi(\boldsymbol{y})}{\sum_{\theta \in S_K} \pi(\boldsymbol{y_\theta})}.$$

Then the transition kernel of $\bar{\boldsymbol{Y}}^\infty$ is

$$P(\bar{\boldsymbol{Y}}^\infty(j+1) \in d\boldsymbol{z} | \bar{\boldsymbol{Y}}^\infty(j) = \boldsymbol{y}) = \sum_{\sigma \in S_K} w(\boldsymbol{y}_\sigma) \alpha_1(y_{\sigma(1)}, dz_{\sigma(1)}) \cdots \alpha_K(y_{\sigma(K)}, dz_{\sigma(K)}).$$

The discrete time numerical approximation to the invariant distribution is

$$\eta_J^\infty \doteq \frac{1}{J} \sum_{j=0}^{J-1} \sum_{\sigma \in S_K} w(\bar{\boldsymbol{Y}}_\sigma^\infty(j)) \delta_{\bar{\boldsymbol{Y}}_\sigma^\infty(j)}.$$

REMARK 5.2. It is not difficult to derive large deviation principles for the discrete time finite swapping or infinite swapping models. However, it remains an open question whether the rate function is monotonic with respect to the swap rate (frequency). However, the discrete time large deviation rate function can be obtained from that of the continuous time pure jump Markov process models through the contraction principle, and the two coincide in the limit as the transition kernels α_i correspond to an infinitesimal time step for the diffusion process (1.1). Hence the discrete time rate function will be at least approximately monotone, and in this sense the infinite swapping limit should (at least approximately) dominate all finite swapping algorithms. This is supported by the data presented in Section 7 and the much more extensive empirical study presented in [14].

6. Partial infinite swapping. As noted in Section 3, the number of weights and their calculation can become unwieldy for infinite swapping even when the number of temperatures is moderate. In this section we construct algorithms that maintain most of the benefit of the infinite swapping algorithm but at a much lower computational cost. In the first subsection we describe the infinite swapping limit models when only a subgroup of the permutations of the particles (respectively, temperatures) are allowed by the prelimit particle (respectively, temperature) swapped process. The computational complexity of these limit models will be controlled by limiting the number of permutations that communicate with each other through the swapping mechanism. The infinite swapping models in this subsection will be called *partial infinite swapping models,* as opposed to the full infinite swapping models in the previous section. The second subsection shows how such partial infinite swapping schemes can be interwoven to approximate the full infinite swapping model.

6.1. Partial infinite swapping models. We consider subsets A of S_K with the property that A is an algebraic subgroup of S_K. That is,

1. the identity belongs to A;
2. if $\sigma_1, \sigma_2 \in A$ then $\sigma_1 \circ \sigma_2 \in A$, where $\circ$ denotes composition;
3. if $\sigma \in A$ then $\sigma^{-1} \in A$.

Although one can write down a partial infinite swapping model that corresponds to instantaneous equilibration for an arbitrary subset A, it is only when A is a subgroup that the corresponding partial infinite swapping process has an interpretation

as the limit of parallel tempering type processes. When alternating between partial infinite swapping processes, a "handoff" rule will be needed, and it is only for those which correspond to subgroups that such a handoff rule is well defined. This point is discussed in some detail in the next section.

The definition of the partial infinite swapping process based on A is completely analogous to that of the full infinite swapping process. The state process $\{\bar{\boldsymbol{Y}}(j) : j = 0, 1, \ldots\}$ is a Markov chain with the transition kernel

$$\alpha^A(\boldsymbol{y}, d\boldsymbol{z}) \doteq \sum_{\sigma \in A} \tilde{w}^A(\boldsymbol{y}_\sigma)\, \alpha_1(y_{\sigma(1)}, dz_{\sigma(1)}) \cdots \alpha_K(y_{\sigma(K)}, dz_{\sigma(K)}) \tag{6.1}$$

and the weighted empirical measure is

$$\tilde{\eta}_J \doteq \frac{1}{J} \sum_{j=0}^{J-1} \sum_{\sigma \in A} \tilde{w}^A(\bar{\boldsymbol{Y}}_\sigma(j)) \delta_{\bar{\boldsymbol{Y}}_\sigma(j)},$$

where the weight function $\tilde{w}^A$ is defined by

$$\tilde{w}^A(\boldsymbol{y}) \doteq \frac{\pi(\boldsymbol{y})}{\sum_{\theta \in A} \pi(\boldsymbol{y}_\theta)}, \tag{6.2}$$

and satisfies for any $\boldsymbol{y}$

$$\sum_{\sigma \in A} \tilde{w}^A(\boldsymbol{y}_\sigma) = 1.$$

We omit the dependence on both $a = \infty$ and A from the notation. Note that in contrast with the full swapping system, it is only those permutations of $\boldsymbol{y}$ corresponding to $\sigma \in A$ that are balanced according to the invariant distribution in their contributions to $\tilde{\eta}_J$.

To illustrate the construction we present a few examples. With a standard abuse of notation denote the permutation σ such that $\sigma(i) = a_i$ by the form $(a_1, a_2, \ldots, a_K)$. In particular, $(1, 2, \ldots, K)$ is the identity of the group S_K.

EXAMPLE 6.1. *Let $K = 4$ and $A = \{(1,2,3,4), (2,1,3,4)\}$. This corresponds to only allowing swaps between temperatures τ_1 and τ_2 at the prelimit. Define*

$$\tilde{w}(\boldsymbol{y}) = \frac{\pi(y_1, y_2, y_3, y_4)}{\pi(y_1, y_2, y_3, y_4) + \pi(y_2, y_1, y_3, y_4)}$$

The probability transition kernel of the corresponding partial infinite swapping process is given by

$$\begin{aligned} &\tilde{w}(y_1, y_2, y_3, y_4)\, \alpha_1(y_1, dz_1)\alpha_2(y_2, dz_2)\alpha_3(y_3, dz_3)\alpha_4(y_4, dz_4) \\ &\quad + \tilde{w}(y_2, y_1, y_3, y_4)\, \alpha_1(y_2, dz_2)\alpha_2(y_1, dz_1)\alpha_3(y_3, dz_3)\alpha_4(y_4, dz_4) \end{aligned}$$

and the contribution to the weighted empirical measure is

$$\tilde{w}\left(\bar{Y}_1, \bar{Y}_2, \bar{Y}_3, \bar{Y}_4\right) \delta_{\left(\bar{Y}_1, \bar{Y}_2, \bar{Y}_3, \bar{Y}_4\right)} + \tilde{w}\left(\bar{Y}_2, \bar{Y}_1, \bar{Y}_3, \bar{Y}_4\right) \delta_{\left(\bar{Y}_2, \bar{Y}_1, \bar{Y}_3, \bar{Y}_4\right)}.$$

Note that with π_{ij} denoting the marginal invariant distribution on the i-th and j-th components, the weight function can be written as

$$\tilde{w}(\boldsymbol{y}) = \frac{\pi_{12}(y_1, y_2)}{\pi_{12}(y_1, y_2) + \pi_{12}(y_2, y_1)},$$

which is consistent with the weights of the two-temperature model in Theorem 2.3.

EXAMPLE 6.2. *Again take $K = 4$, but this time use the subgroup generated by $(2,1,3,4)$ and $(1,2,4,3)$, i.e., $A = \{(1,2,3,4),(2,1,3,4),(1,2,4,3),(2,1,4,3)\}$. Then the dynamics are given by*

$$\begin{aligned}
&\tilde{w}\left(y_1,y_2,y_3,y_4\right)\alpha_1(y_1,dz_1)\alpha_2(y_2,dz_2)\alpha_3(y_3,dz_3)\alpha_4(y_4,dz_4)\\
&\quad+\tilde{w}\left(y_2,y_1,y_3,y_4\right)\alpha_1(y_2,dz_2)\alpha_2(y_1,dz_1)\alpha_3(y_3,dz_3)\alpha_4(y_4,dz_4)\\
&\qquad+\tilde{w}\left(y_1,y_2,y_4,y_3\right)\alpha_1(y_1,dz_1)\alpha_2(y_2,dz_2)\alpha_3(y_4,dz_4)\alpha_4(y_3,dz_3)\\
&\qquad\quad+\tilde{w}\left(y_2,y_1,y_4,y_3\right)\alpha_1(y_2,dz_2)\alpha_2(y_1,dz_1)\alpha_3(y_4,dz_4)\alpha_4(y_3,dz_3)
\end{aligned}$$

where the weight function $\tilde{w}$ is defined by

$$\widetilde{w}\left(\boldsymbol{y}\right)=\frac{\pi\left(y_1,y_2,y_3,y_4\right)}{\pi\left(y_1,y_2,y_3,y_4\right)+\pi\left(y_2,y_1,y_3,y_4\right)+\pi\left(y_1,y_2,y_4,y_3\right)+\pi\left(y_2,y_1,y_4,y_3\right)}.$$

The contribution to the weighted empirical measure is

$$\begin{aligned}
&\tilde{w}\left(\bar{Y}_1,\bar{Y}_2,\bar{Y}_3,\bar{Y}_4\right)\delta_{\left(\bar{Y}_1,\bar{Y}_2,\bar{Y}_3,\bar{Y}_4\right)}+\tilde{w}\left(\bar{Y}_2,\bar{Y}_1,\bar{Y}_3,\bar{Y}_4\right)\delta_{\left(\bar{Y}_2,\bar{Y}_1,\bar{Y}_3,\bar{Y}_4\right)}\\
&\quad+\tilde{w}\left(\bar{Y}_1,\bar{Y}_2,\bar{Y}_4,\bar{Y}_3\right)\delta_{\left(\bar{Y}_1,\bar{Y}_2,\bar{Y}_4,\bar{Y}_3\right)}+\tilde{w}\left(\bar{Y}_2,\bar{Y}_1,\bar{Y}_4,\bar{Y}_3\right)\delta_{\left(\bar{Y}_2,\bar{Y}_1,\bar{Y}_4,\bar{Y}_3\right)}.
\end{aligned}$$

EXAMPLE 6.3. *We let $K = 3$ and take A to be the subgroup of S_K generated by the rotation $(2,3,1)$, i.e., $A = \{(1,2,3),(2,3,1),(3,1,2)\}$. Then the dynamics are given by*

$$\begin{aligned}
&\tilde{w}\left(y_1,y_2,y_3\right)\alpha_1(y_1,dz_1)\alpha_2(y_2,dz_2)\alpha_3(y_3,dz_3)\\
&\quad+\tilde{w}\left(y_2,y_3,y_1\right)\alpha_1(y_2,dz_2)\alpha_2(y_3,dz_3)\alpha_3(y_1,dz_1)\\
&\qquad+\tilde{w}\left(y_3,y_1,y_2\right)\alpha_1(y_3,dz_3)\alpha_2(y_1,dz_1)\alpha_3(y_2,dz_2)
\end{aligned}$$

where

$$\tilde{w}\left(\boldsymbol{y}\right)=\frac{\pi(y_1,y_2,y_3)}{\pi(y_1,y_2,y_3)+\pi\left(y_2,y_3,y_1\right)+\pi\left(y_3,y_1,y_2\right)}$$

and the contribution to the weighted empirical measure is

$$\tilde{w}_1\left(\bar{Y}_1,\bar{Y}_2,\bar{Y}_3\right)\delta_{\left(\bar{Y}_1,\bar{Y}_2,\bar{Y}_3\right)}+\tilde{w}_2(\bar{Y}_2,\bar{Y}_3,\bar{Y}_1)\delta_{\left(\bar{Y}_2,\bar{Y}_3,\bar{Y}_1\right)}+\tilde{w}_3\left(\bar{Y}_3,\bar{Y}_1,\bar{Y}_3\right)\delta_{\left(\bar{Y}_3,\bar{Y}_1,\bar{Y}_3\right)}.$$

The first two examples would correspond to the infinite swapping limit of a standard parallel tempering process, where swaps between only 1 and 2 are allowed in the first example, and swaps between 1 and 2 and swaps between 3 and 4 are allowed in the second. Note that the computational complexity does not increase significantly between the first and second example. The third example corresponds to a very different sort of prelimit process, in which "rotations" of the coordinates $(y_1,y_2,y_3)\to(y_2,y_3,y_1)\to(y_3,y_1,y_2)\to(y_1,y_2,y_3)$ are allowed. One can devise a Metropolis type rule that allows such "swaps" and yields the indicated infinite swapping system.

6.2. Approximating full infinite swapping by partial swapping. In this section we consider the issue of alternating between such partial infinite swapping systems to approximate the full infinite swapping limit. Let A and B be subgroups of S_K. A and B are said to *generate* S_K if the smallest subgroup that contains A and B

is S_K itself. Note that the total number of permutations in $A \cup B$ can be significantly smaller than $K!$, the size of S_K. In fact, it is possible to construct subgroups A and B that generate S_K and that the total number of permutations in $A \cup B$ is of order K. There is an obvious extension to more than two subgroups.

EXAMPLE 6.4. *Let $K = 4$ and let A be generated by $\{(2,1,3,4),(1,3,2,4)\}$ and B be generated by $\{(1,3,2,4),(1,2,4,3)\}$, respectively. Thus A is the collection of 6 permutations that fix the last component and allow all rearrangements of the first three, while B fixes the first component and allows all rearrangements of the last three. Then A and B generate S_K.*

EXAMPLE 6.5. *Let $K = 4$ and let A and B be subgroups generated by $\{(2,1,3,4)\}$ and $\{(2,3,4,1)\}$, respectively. In other words, $A = \{(1,2,3,4),(2,1,3,4)\}$ corresponds to only allowing permutations between the first two components, while $B = \{(1,2,3,4),(2,3,4,1),(3,4,1,2),(4,1,2,3)\}$ corresponds to cycling of the four temperatures. Then A and B generate S_K.*

To keep the computational cost controlled, one can approximate the full infinite swapping model by alternating between partial infinite swapping processes whose associated subgroups generate the whole group. However, one must be careful in how the "handoff" is made when switching between different partial swapping models. It turns out that one cannot simply switch between different partial infinite swapping dynamics (i.e., transition kernels). Recall that in order to get a consistent approximation to the desired target invariant distribution we do not use the empirical measure generated by $\bar{\boldsymbol{Y}}$, but rather a carefully constructed weighted empirical measure that works with several permutations of $\bar{\boldsymbol{Y}}$. Simply switching the dynamics and weights will in fact produce an algorithm that may not converge to the target distribution.

To see how one should design a handoff rule, note that if one considers a collection of transition kernels each having the same invariant distribution and alternates between them in a way that does not depend on the outcomes prior to a switch, then the resulting empirical measure will in fact converge to the common invariant distribution. This fact is used (at least implicitly) in the parallel tempering algorithm itself, where one alternates the pair of particles being considered for swapping according to deterministic or random rules so long as the random rules do not rely on previously observed outcomes.

Now we use the fact that each partial infinite swapping model is a limit of either a parallel tempering algorithm where only some pairs of particles are considered for swapping, or some more general form of parallel tempering which would allow groups of particles to simultaneously swap (according to an appropriate Metropolis-type acceptance rule). An example in the earlier category would be A in Example 6.4, which arises if only the pairs corresponding to temperatures τ_1, τ_2 and τ_2, τ_3 are allowed to swap, whereas an example in the latter category would be B in Example 6.5, which corresponds to allowing the particle at temperature τ_i to move to the location of the particle at temperature τ_{i-1} (with $\tau_0 = \tau_4$), and the reverse. Furthermore, each of these partial infinite swapping models arises as a limit of transition kernels of the corresponding temperature swapped processes which preserve the same common invariant distribution. In taking the limit as the swap rate tends to infinity, the correspondence between particle locations for a particle swapped process and the "instantaneously equilibrated" temperature swapped process $\bar{Y}$ is lost. However, one can construct a consistent algorithm by *reconstructing this correspondence*. In fact one should choose the particle location according to the probabilities (under the invariant distribution) associated with the various permutations in the subgroup. See Subsection 6.3 for

more detailed discussion on the intuition behind this approximation algorithm.

We next present an algorithm for alternating between two partial infinite swapping dynamics. The restriction to two is for notational convenience only. Suppose that the dynamics are indexed by the corresponding subgroups A and B, and that n_A steps of subgroup A are to be alternated with n_B steps of subgroup B. For simplicity we do not describe a "burn-in" period. As in (6.1) and (6.2) we let $\alpha^A(\boldsymbol{y}, d\boldsymbol{z})$ and $\tilde{w}^A(\boldsymbol{y})$ denote the transition kernel for A and the weights allocated to the permutation $\sigma \in A$, respectively, and similarly for B.

ALGORITHM 6.6. *(Approximation to full infinite swapping)*

1. *Initialization:* $\bar{\boldsymbol{X}}^A(0) = \bar{\boldsymbol{Y}}(0;0) \in (\mathbb{R}^d)^K, \ell = 1$.
2. *Loop ℓ:*
   (a) *Initialization for A dynamics: set* $\bar{\boldsymbol{Y}}(\ell;0) = \bar{\boldsymbol{X}}^A(\ell-1)$.
   (b) *Subgroup A dynamics: update* $\bar{\boldsymbol{Y}}(\ell;k), k = 1, \ldots, n_A$ *according to the transition kernel* α^A*, and add*
   $$\sum_{\sigma \in A} \tilde{w}^A(\bar{\boldsymbol{Y}}_\sigma(\ell;k))\delta_{\bar{\boldsymbol{Y}}_{\boldsymbol{\sigma}}(\ell;k)}$$
   to the un-normalized empirical measure.
   (c) *Reconstructing particle locations at the end of A dynamics: Let* $\bar{\boldsymbol{X}}^B(\ell)$ *be a random sample from the set* $\{\bar{\boldsymbol{Y}}_{\boldsymbol{\sigma}}(\ell;n_A) : \sigma \in A\}$ *according to the weights* $\{\tilde{w}^A(\bar{\boldsymbol{Y}}_\sigma(\ell;n_A)) : \sigma \in A\}$.
   (d) *Initialization for B dynamics: set* $\bar{\boldsymbol{Y}}(\ell;n_A) = \bar{\boldsymbol{X}}^B(\ell)$.
   (e) *Subgroup B dynamics: update* $\bar{\boldsymbol{Y}}(\ell;k), k = n_A+1, \ldots, n_A+n_B$ *according to the transition kernel* α^B*, and add*
   $$\sum_{\sigma \in B} \tilde{w}^B(\bar{\boldsymbol{Y}}_\sigma(\ell;k))\delta_{\bar{\boldsymbol{Y}}_{\boldsymbol{\sigma}}(\ell;k)}$$
   to the un-normalized empirical measure.
   (f) *Reconstructing particle locations at the end of B dynamics: Let* $\bar{\boldsymbol{X}}^A(\ell)$ *be a random sample from the set* $\{\bar{\boldsymbol{Y}}_{\boldsymbol{\sigma}}(\ell;n_A+n_B) : \sigma \in B\}$ *according to the weights* $\{\tilde{w}^B(\bar{\boldsymbol{Y}}_\sigma(\ell;n_A+n_B)) : \sigma \in B\}$.
   (g) *Set $\ell = \ell+1$ and loop back to (a).*
3. *Normalize the empirical measure.*

6.3. Discussions on the approximation. In this section we further discuss the intuition underlying the handoff rule between different partial infinite swapping dynamics and the approximation algorithm of the previous section. We temporarily assume that the model is in continuous time since the intuition is most transparent in this case.

For simplicity let us assume that there are three temperatures and two groups $A = \{(1,2,3),(2,1,3)\}$ and $B = \{(1,2,3),(1,3,2)\}$. That is, under group A dynamics only pairwise swaps between the temperatures τ_1 and τ_2 are allowed, while under the group B dynamics, only the swaps between τ_2 and τ_3 are allowed. See Figure 6.1.

Consider the following prelimit swapping model. Let the swap rate be a. The dynamics corresponding to group A and group B will be alternated on time intervals of length h. Hence on the interval $[2kh,(2k+1)h)$ the particle swapped process $(\bar{X}_1^a, \bar{X}_2^a, \bar{X}_3^a)$ only involves swaps between temperatures τ_1 and τ_2. One can easily construct the corresponding temperature swapped process $(\bar{Y}_1^a, \bar{Y}_2^a, \bar{Y}_3^a)$ as before. Note that $\bar{X}_3^a = \bar{Y}_3^a$ on this time interval. Similarly, on the interval $[(2k+1)h,(2k+$

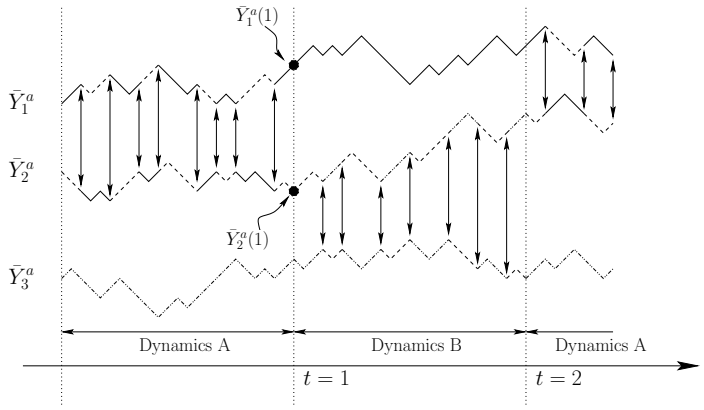

FIG. 6.1. *Approximation via partial infinite swapping*

$2)h$), only swaps between τ_2 and τ_3 are allowed and on this interval $\bar{X}_1^a = \bar{Y}_1^a$. Note that there is no ambiguity for the prelimit processes at the switch times $t = h, 2h, \ldots$, since the locations of the particles $(\bar{X}_1^a, \bar{X}_2^a, \bar{X}_3^a)$ are known.

Now consider the limit as $a \to \infty$ with h being fixed. Without loss of generality, we will only discuss how to deal with the switch of the dynamics at time $t = h$. On the time interval $[0, h)$ we have the partial infinite swapping limit process $\bar{\boldsymbol{Y}}^A = (\bar{Y}_1, \bar{Y}_2, \bar{Y}_3)$ that corresponds to the group A. Similarly it is clear that on the time interval $[h, 2h)$ we should have the partial infinite swapping process $\bar{\boldsymbol{Y}}^B$ corresponding to the group B. The problem is, however, by taking the limit, we lose the information on the locations of the particles $(\bar{X}_1^a, \bar{X}_2^a, \bar{X}_3^a)$. Unless we can somehow recover this information at the switch time $t = h$ to assign $\bar{\boldsymbol{Y}}^B(h)$, we cannot determine the dynamics of $\bar{\boldsymbol{Y}}^B$ on $[h, 2h)$. The key is to recall that the infinite swapping limit instantaneously equilibrates multiple locations according to the invariant distribution. In other words, given $\bar{\boldsymbol{Y}}^A(h-) = \boldsymbol{y} = (y_1, y_2, y_3)$, the locations of the particles are distributed according to

$$\sum_{\sigma \in A} \tilde{w}^A(\boldsymbol{y}_\sigma)\delta_{\boldsymbol{y}_\sigma} = \frac{\pi(y_1, y_2, y_3)\delta_{(y_1,y_2,y_3)} + \pi(y_2, y_1, y_3)\delta_{(y_2,y_1,y_3)}}{\pi(y_1, y_2, y_3) + \pi(y_2, y_1, y_3)}.$$

Therefore, in order to identify the locations of the particles at time h, we will take a random sample from this distribution once $\bar{\boldsymbol{Y}}^A(h-)$ is known. This explains the handoff rule used at the switch times of partial infinite swapping processes.

Now we let $h \to 0$. Since A and B generate the whole permutation group S_K, it is easy to check that at each time instant, the locations of $\{\boldsymbol{y}_\sigma : \sigma \in S_K\}$ are equilibrated according to their invariant distribution, and therefore in the limit we will attain the full infinite swapping model. This can be made rigorous by exploiting the time scale separation between the slow diffusion processes $(\bar{\boldsymbol{Y}}^A, \bar{\boldsymbol{Y}}^B)$ and the fast switching process. We omit the proof because the discussion is largely motivational.

Coming back to the discrete time partial infinite swapping model, it is clear that Algorithm 6.6 is nothing but a straightforward adaption of the preceding discussion to discrete time. The only difference is that one cannot establish an analogous result regarding approximation to the full infinite swapping model as in continuous time. The subtlety here is that in continuous time, as $h \to 0$, one can basically ignore any

effect from the diffusion on any small time interval and assume that the process is only making jumps between different permutations of a fixed triple (y_1, y_2, y_3). This time scale separation is no longer valid in discrete time. In this setting, the performance of a scheme based on interweaving partial infinite swapping schemes lies between parallel tempering and full infinite swapping, and computational results suggest that it is closer to the latter than the former.

The issue of which interwoven partial schemes will perform best is an open question. In practice we have used schemes of the following form. Suppose that a set of say 45 temperatures is given. We then partition 45 into blocks of sizes $3, 6, \ldots, 6$, with the first block containing the lowest three temperatures, the second block the next six, and so on. Dynamic A then is given by allowing all permutations within each block. Note that the complexity of the coefficients is then no worse than 6!. In Dynamic B we use the partition $6, 6, \ldots, 6, 3$. The form of the partial scheme is heuristically motivated by allowing the largest possible overlap between the different blocks when switching between dynamics, subject to the constraint that blocks be of size no greater than 6.

7. Numerical examples. In this section we present data comparing parallel tempering at various swap rates and both full and partial infinite swapping. We present what we call "relaxation studies." The quantity of interest is the average potential energy of the lowest temperature component under the invariant distribution. In these studies, the system is run a long time to reach equilibrium, after which it is repeatedly pushed out of equilibrium and we measure the time needed to "relax" back to equilibrium. Each cycle consists of temporarily raising the temperatures of some of the lowest temperature components for a number of steps sufficient to push the average potential energy away from the "true" value (as measured by either sample or time averages). The temperatures are then returned to their "true" values for a fixed number of steps, and the process is then repeated 2,000 times. We plot the average of the 2,000 samples as a function of the number of moves, and the performance of the algorithm is captured by the rate at which these averages approach the correct value.

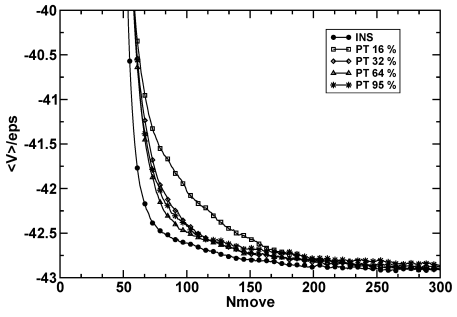

Figure 3

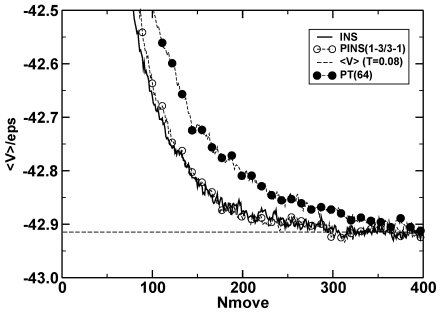

Figure 4

Figures 3 and 4 present data for a Lennard-Jones cluster of 13 atoms, using the "smart Monte Carlo" scheme of [18] for the simulation of the dynamics, which produces a relatively large move in configuration space for each step. The "true" value is approximately -42.92. This is a relatively simple model, and was studied using only

4 temperatures. The temperatures are dropped to the true values at step 50. Infinite swapping converges more rapidly than any of the parallel tempering schemes. We see in Figure 3 that the most efficient of the parallel tempering schemes appears to use an attempted swap rate of around 64%. [The rates that would typically be used in such calculations are in the range of 5-10%.] Figure 4 magnifies a portion of the graph, but plots only the best parallel tempering result and adds a partial infinite swapping result based on blocks of the form 1,3 and 3,1, and with a handoff at each Metropolis step. Little difference is observed between the partial and full forms, though exclusive use of either of the partial forms by itself performs poorly.

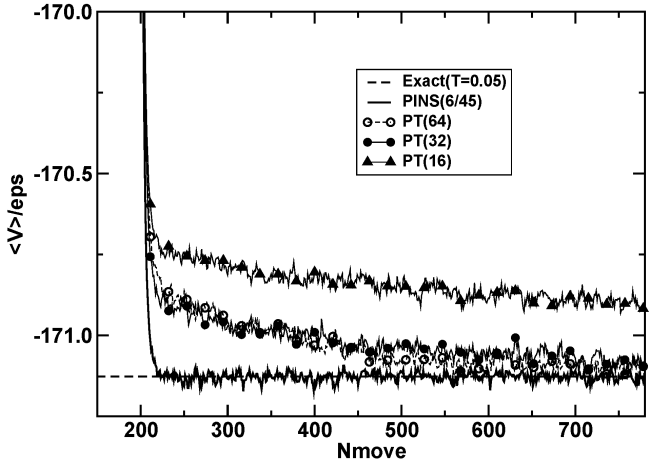

Figure 5

The Lennard-Jones cluster of 13 atoms is not a particularly demanding problem, but is presented so a comparison can be made between the full and partial infinite swapping forms. A much more complex example is the Lennard-Jones cluster of 38 atoms. Data for this example obtained using a 45 temperature ensemble is given in Figure 5. Because full infinite swapping is impossible for this larger computational ensemble, we use the partial form. For comparison, results are also presented for parallel tempering.

The details concerning the computational methods underlying both the parallel tempering and infinite swapping results of Figure 5 along with a discussion of the temperature ensemble involved for this example can be found in [14]. Briefly summarized, as with the previous 13-atom Lennard-Jones example, Figure 5 denotes the results of a series of relaxation experiments. Here, however, 45 temperatures are used with the lowest 15 being involved in the heating/cooling process. The heating and cooling cycles consist of 1200 smart Monte Carlo moves, each of one unit Lennard-Jones time duration. The cooling segment is taken as the portion of the cycle from moves 200 to 800 with the remainder being the heating portion. During the cooling portion of the cycle the 45 temperatures in the ensemble cover the range from (0.050-0.210) in temperature steps of 0.005, and from (0.210 - 0.330) in steps of 0.010 while during the heating portion of the cycle temperatures less than or equal to 0.150 are set equal to 0.150. The results shown in Figure 5 are obtained using 600 thermal cycles.

The 38-atom Lennard-Jones cluster has an interesting landscape. In particular,

while the global and lowest-lying local minima are similar in energy, the minimum energy pathway that separates them involves appreciably higher energies and contains 13 separate barriers [24, Chapter 8.3]. As discussed in [14], the partial infinite swapping approach is appreciably more effective than conventional tempering approaches in providing a proper sampling of this complex potential energy landscape.

8. Appendix.

8.1. Proof of Theorem 4.1. Throughout the proof, we let $S = S_1 \times S_1$, where $S_1 \subset \mathbb{R}^d$ is convex and compact. Let $\mathcal{P}(S)$ denote the Polish space of all probability measures on S equipped with the topology of weak convergence. For any probability measure $\nu \in \mathcal{P}(S)$, define its mirror image $\nu^R \in \mathcal{P}(S)$ by requiring

$$\nu^R(A \times B) = \nu(B \times A)$$

for all Borel sets $A, B \subset \mathbb{R}^d$. Furthermore, as in the rest of the paper, a bold symbol $\boldsymbol{x} \in S$ means $\boldsymbol{x} = (x_1, x_2)$, where $x_1, x_2 \in \mathbb{R}^d$, and $\boldsymbol{x}^R = (x_2, x_1)$. We also use the notation

$$\alpha(\boldsymbol{x}, d\boldsymbol{y}) \doteq \alpha_1(x_1, dy_1)\alpha_2(x_2, dy_2),$$

which is a probability transition kernel defined on S given S.

To prove the uniform large deviation principle, it suffices to prove the equivalent uniform Laplace principle [3, Chapter 1]. To simplify the proof we have assumed that S is compact. This would be the case if, e.g., V is defined with periodic boundary conditions. The general case can be handled under (2.1) by using V as a Lyapunov function [3, Section 8.2]. It will be convenient to split this into upper and lower bounds. We also consider just the (more complicated) case where $a_T \to \infty$ but $a_T < \infty$ for each T. Allowing $a_T = \infty$ requires a different notation to handle this special case, but does not change the structure of the proof otherwise.

We will show for any bounded continuous function $F : \mathcal{P}(S) \to \mathbb{R}$ that

$$\lim_{T\to\infty} -\frac{1}{T} \log E\left[\exp\{-TF(\eta_T^{a_T})\}\right] = \inf_{\nu\in\mathcal{P}(S)} \left[F(\nu) + I^\infty(\nu)\right]. \tag{8.1}$$

By adding a constant to both sides we can assume $F \geq 0$, and do so for the rest of this section.

The proof of the uniform large deviation principle is based on the weak convergence approach. The proof is complicated by the multiscale aspect of the fast swapping process, as well as the fact that $\eta_T^{a_T}$ is a weighted empirical measure that involves this fast process.

8.2. Preliminary results.

8.2.1. A representation. We first state a stochastic control representation for the left hand side of (8.1). As with the derivation of the infinite swapping process via weak convergence, it will be necessary to work with the (distributionally equivalent) temperature swapped processes for tightness to hold. In the representation, all random variables used to construct $\eta_T^{a_T}$ are replaced by random variables whose distribution is selected, and both the distributions and the random variables will be distinguished from their uncontrolled, original counterparts by an overbar. For this reason, while the continuous time process is denoted by $\boldsymbol{Y}^a(t)$, we change notation and use $\boldsymbol{U}^a(j)$ rather than $\bar{\boldsymbol{Y}}^a(j)$ to denote the discrete time process.

We first construct the temperature swapped process. Let $\alpha(\boldsymbol{x}, d\boldsymbol{y}|0) = \alpha(\boldsymbol{x}, d\boldsymbol{y})$ and $\alpha(\boldsymbol{x}, d\boldsymbol{y}|1) = \alpha(\boldsymbol{x}^R, d\boldsymbol{y}^R)$, and let $\{N_j^a, j = 0, 1, \ldots\}$ be iid geometric random variables with parameter $1/(1+a)$, i.e. geometric random variables with mean a. Then the random variables $\{\boldsymbol{U}^a(j), j = 0, 1, \ldots\}$, $\{M_\ell^a(j), j = 0, 1, \ldots, \ell = 0, 1, \ldots, N_j^a\}$ are constructed recursively as follows. Given $M_0^a(j) = z$ and $\boldsymbol{U}^a(j) = \boldsymbol{x}$, $\boldsymbol{U}^a(j+1)$ is distributed according to $\alpha(\boldsymbol{x}, d\boldsymbol{y}|z)$. The process $M_\ell^a(j), \ell = 0, 1, \ldots, N_j^a$ is a Markov chain with states $\{0, 1\}$ and transition probabilities

$$\begin{array}{cc} p(0,0|\boldsymbol{x}) = g(\boldsymbol{x}) & p(0,1|\boldsymbol{x}) = 1 - g(\boldsymbol{x}) \\ p(1,0|\boldsymbol{x}) = 1 - g(\boldsymbol{x}^R) & p(1,0|\boldsymbol{x}) = g(\boldsymbol{x}^R) \end{array}. \tag{8.2}$$

The initial value for the subsequent interval is given by $M_0^a(j+1) = M_{N_j^a}^a(j)$. Letting $\{\tau_{j,\ell}^a, i = 0, 1, \ldots, \ell = 0, 1, \ldots, N_j^a - 1\}$ be iid exponential random variables with mean $1/a$, the temperature swapped process in continuous time is then given by

$$\boldsymbol{Y}^a(t) = \boldsymbol{U}^a(j) \text{ for } \sum_{i=0}^{j-1} \sum_{\ell=0}^{N_j^a - 1} \tau_{i,\ell}^a \le t < \sum_{i=0}^{j} \sum_{\ell=0}^{N_j^a - 1} \tau_{i,\ell}^a,$$

(with the convention that the sum from 0 to -1 is 0),

$$Z^a(t) = M_\ell^a(j) \text{ for } \sum_{i=0}^{j-1} \sum_{k=0}^{\ell-1} \tau_{i,k}^a \le t < \sum_{i=0}^{j-1} \sum_{k=0}^{\ell} \tau_{i,k}^a,$$

and lastly the ordinary and weighted empirical measures are given by

$$\psi_T^a = \frac{1}{T} \int_0^T \delta_{\boldsymbol{Y}^a(t)} dt \text{ and } \eta_T^a = \frac{1}{T} \int_0^T \left[1_{\{Z^a(t)=0\}} \delta_{\boldsymbol{Y}^a(t)} + 1_{\{Z^a(t)=1\}} \delta_{\boldsymbol{Y}^a(t)^R} \right] dt.$$

Let σ^a denote the exponential distribution with mean $1/a$ and let β^a denote the geometric distribution with mean a. For the representation, all distributions [e.g., $\alpha(\boldsymbol{x}, d\boldsymbol{y}|z)$], can be perturbed from their original form, but such a perturbation pays a relative entropy cost. We distinguish the new distributions and random variables by using an overbar. Given $T \in (0, \infty)$, let R^a and K^a be the discrete time indices when the continuous time parameter reaches T, i.e.,

$$\sum_{i=0}^{R^a-2} \sum_{k=0}^{N_i^a - 1} \tau_{i,k}^a + \sum_{k=0}^{K^a-1} \tau_{R^a-1,k}^a \le T < \sum_{i=0}^{R^a-2} \sum_{k=0}^{N_i^a - 1} \tau_{i,k}^a + \sum_{k=0}^{K^a} \tau_{R^a-1,k}^a. \tag{8.3}$$

In this representation the barred quantities are constructed analogously to their unbarred counterparts. Thus, e.g., $\bar{R}^a$ and $\bar{N}_i^a$ are defined by (8.3) but with $\tau_{i,k}^a$ replaced by $\bar{\tau}_{i,k}^a$. Random variables corresponding to any given value of j are constructed in the order $\bar{\boldsymbol{U}}^a(j+1), \bar{N}_j^a, \bar{M}_\ell^a(j), \bar{\tau}_{i,\ell}^a, \ell = 0, 1, \ldots, \bar{N}_j^a$, and then j is updated to $j+1$. Barred measures, which are also allowed to depend on discrete time, are used to construct the corresponding barred random variables, e.g., $\bar{\boldsymbol{U}}^a(j+1)$ is (conditionally) distributed according to $\bar{\alpha}_j(\bar{\boldsymbol{U}}^a(j), \cdot|\bar{M}_0^a(j))$. The infimum is over all collections of measures $\{\bar{\alpha}_j, \bar{\beta}_j^a, \bar{p}_{j,\ell}, \bar{\sigma}_{j,\ell}^a\}$ and, although this is not denoted explicitly, any particular measure can depend on all previously constructed random variables. To simplify

notation we let $\bar{N}^a_{\bar{R}^a}$ denote $\bar{K}^a$. We state the representation for $\{\eta^a_T\}$, and note that an analogous representation holds for $\{\psi^a_T\}$.

LEMMA 8.1. *Let $G : \mathcal{P}(S) \times \mathbb{N} \to \mathbb{R}$ be bounded from below and measurable. Then the representation*

$$-\frac{1}{T}\log E\left[\exp\{-TG(\eta^a_T, R^a)\}\right] = \inf E\left[G(\bar{\eta}^a_T, \bar{R}^a) \right.$$
$$+ \frac{1}{T}\sum_{i=0}^{\bar{R}^a-1}\left[R\left(\bar{\alpha}_i(\bar{\boldsymbol{U}}^a(i),\cdot|\bar{M}^a_0(i))\,\middle\|\,\alpha(\bar{\boldsymbol{U}}^a(i),\cdot|\bar{M}^a_0(i))\right) + R\left(\bar{\beta}^a_i\,\|\beta^a\right)\right]$$
$$\left. + \frac{1}{T}\sum_{i=0}^{\bar{R}^a-1}\sum_{k=0}^{\bar{N}^a_i-1}\left[R\left(\bar{p}_{i,k}(\bar{M}^a_k(i),\cdot|\bar{\boldsymbol{U}}^a(i))\,\middle\|\,p(\bar{M}^a_k(i),\cdot|\bar{\boldsymbol{U}}^a(i))\right) + R\left(\bar{\sigma}^a_{i,k}\,\|\sigma^a\right)\right] \right]$$

is valid.

The proof of such representations follow from the chain rule for relative entropy (see, e.g., [3, Section B.2]). A novel feature of the representation here is that the total number of discrete time steps is random. However, this case can easily be reduced to the case with a fixed deterministic number of steps.

8.2.2. Rate for the ordinary empirical measure. Notation for marginals. We will frequently factor measures on product spaces in the proof. For a (deterministic) probability measure ν on a product space such as $S^1 \times S^2 \times S^3$, with each S^i a Polish space, we use notation such as $\nu_{1,2}$ to denote the marginal distribution on the first 2 components, and notation such as $\nu_{1|3}$ to denote the conditional distribution on the first component given the third. When ν is a measurable random measure these can all be chosen so that they are also measurable.

We will make use of the rate function for the ordinary empirical measure. Let $\varphi(\boldsymbol{x}, d\boldsymbol{y}) = \alpha(\boldsymbol{x}, d\boldsymbol{y}|0)\rho_0(\boldsymbol{x}) + \alpha(\boldsymbol{x}, d\boldsymbol{y}|1)\rho_1(\boldsymbol{x})$, where $\rho_0(\boldsymbol{x}) = \rho(\boldsymbol{x})$ and $\rho_1(\boldsymbol{x}) = \rho(\boldsymbol{x}^R)$, and let $\bar{\mu}(d\boldsymbol{x}) = [\pi(\boldsymbol{x}) + \pi(\boldsymbol{x}^R)]d\boldsymbol{x}/2$ be its unique invariant probability distribution. If γ is absolutely continuous with respect to $\bar{\mu}$ with $\kappa(\boldsymbol{x}) = [d\gamma/d\bar{\mu}](\boldsymbol{x})$, then set

$$K(\gamma) = 1 - \int \sqrt{\kappa(\boldsymbol{x})\kappa(\boldsymbol{y})}\bar{\mu}(d\boldsymbol{x})\varphi(\boldsymbol{x}, d\boldsymbol{y}).$$

Note that K is convex. We then extend the definition to all of $\mathcal{P}(S)$ via lower semicontinuous regularization with respect to the weak topology. Thus if $\gamma_i \to \gamma$ in the weak topology and if each γ_i is absolutely continuous with respect to $\bar{\mu}$, then $\liminf_i K(\gamma_i) \geq K(\gamma)$, and we have equality for at least one such sequence. Note that since $\bar{\mu}$ is mutually absolutely continuous with respect to Lebesgue measure, this means that $K(\gamma) \leq 1$ for all $\gamma \in \mathcal{P}(S)$.

The following lemma will help relate weak limits of quantities in the representations to the rate function of the ordinary empirical measure.

LEMMA 8.2. *Let $\gamma \in \mathcal{P}(S)$ be absolutely continuous with respect to $\bar{\mu}$ and let $\kappa = [d\gamma/d\bar{\mu}]$. Assume $A \in (0,\infty), \nu \in \mathcal{P}(S\times S)$ is such that $[\nu]_1 = [\nu]_2$, $R(\nu\|[\nu]_1\otimes\varphi) < \infty$,*

r is such that $r[\nu]_1 = \kappa\bar{\mu}$, and $0 \leq -\int \log r(\boldsymbol{y})[\nu]_1(d\boldsymbol{y}) < \infty$. Then

$$K(\gamma) = 1 - \int \sqrt{\kappa(\boldsymbol{x})\kappa(\boldsymbol{y})}\bar{\mu}(d\boldsymbol{x})\varphi(\boldsymbol{x}, d\boldsymbol{y}) \tag{8.4}$$
$$\leq AR(\nu\|[\nu]_1 \otimes \varphi) - A\int \log r(\boldsymbol{y})[\nu]_1(d\boldsymbol{y}) + A\log A - A + 1.$$

Proof. Let C be the set where $\kappa(\boldsymbol{x}) = 0$. Then $-\int \log r(\boldsymbol{y})[\nu]_1(d\boldsymbol{y}) < \infty$ implies $r(\boldsymbol{y}) > 0$ a.s. with respect to $[\nu]_1(d\boldsymbol{y})$, and we also have $\int r(\boldsymbol{y})[\nu]_1(d\boldsymbol{y}) = 1$, so that $r(\boldsymbol{y}) < \infty$ a.s. with respect to $[\nu]_1(d\boldsymbol{y})$. It follows that $[\nu]_1(C) = 0$. Now suppose that

$$\int \sqrt{\kappa(\boldsymbol{x})\kappa(\boldsymbol{y})}\bar{\mu}(d\boldsymbol{x})\varphi(\boldsymbol{x}, d\boldsymbol{y}) = 0.$$

Then $\kappa(\boldsymbol{x})\kappa(\boldsymbol{y}) = 0$ a.s. with respect to Lebesgue measure, and so if $\kappa(\boldsymbol{x}) \neq 0$ then $\varphi(\boldsymbol{x}, \{\boldsymbol{y} : \kappa(\boldsymbol{y}) > 0\}) = 0$, or $\varphi(\boldsymbol{x}, C) = 1$. Thus $\nu((S\backslash C) \times C) = 0$, while $[\nu]_1 \otimes \varphi((S\backslash C) \times C) = 1$, which implies $R(\nu\|[\nu]_1 \otimes \varphi) = \infty$, which is a contradiction. We conclude that

$$\int \sqrt{\kappa(\boldsymbol{x})\kappa(\boldsymbol{y})}\bar{\mu}(d\boldsymbol{x})\varphi(\boldsymbol{x}, d\boldsymbol{y}) > 0.$$

Since also

$$\int \sqrt{\kappa(\boldsymbol{x})\kappa(\boldsymbol{y})}\bar{\mu}(d\boldsymbol{x})\varphi(\boldsymbol{x}, d\boldsymbol{y}) \leq 1,$$

it follows that

$$-\log \int \sqrt{\kappa(\boldsymbol{x})\kappa(\boldsymbol{y})}\bar{\mu}(d\boldsymbol{x})\varphi(\boldsymbol{x}, d\boldsymbol{y}) \in [0, \infty).$$

Since $R(\nu\|[\nu]_1 \otimes \varphi) < \infty$ and since $\bar{\mu}$ is the invariant distribution of φ it follows that $R([\nu]_1\|\bar{\mu}) < \infty$ [3, Lemma 8.6.2]. This means that

$$\int \log \frac{\kappa(\boldsymbol{x})}{r(\boldsymbol{x})}[\nu]_1(d\boldsymbol{x}) < \infty,$$

and since $-\int \log r(\boldsymbol{y})[\nu]_1(d\boldsymbol{y}) < \infty$, it follows that $-\int \log \kappa(\boldsymbol{x})[\nu]_1(d\boldsymbol{x}) > -\infty$. From $[\nu]_1 = [\nu]_2$ we conclude that

$$-\frac{1}{2}\int [\log \kappa(\boldsymbol{x}) + \log \kappa(\boldsymbol{y})]\, \nu(d\boldsymbol{x}, d\boldsymbol{y}) > -\infty. \tag{8.5}$$

By relative entropy duality ([3, Proposition 1.4.2])

$$-\log \int \sqrt{\kappa(\boldsymbol{x})\kappa(\boldsymbol{y})}\bar{\mu}(d\boldsymbol{x})\varphi(\boldsymbol{x}, d\boldsymbol{y})$$
$$= -\log \int e^{\frac{1}{2}[\log \kappa(\boldsymbol{x}) + \log \kappa(\boldsymbol{y})]}\bar{\mu}(d\boldsymbol{x})\varphi(\boldsymbol{x}, d\boldsymbol{y})$$
$$\leq R(\nu\|\bar{\mu} \otimes \varphi) - \frac{1}{2}\int [\log \kappa(\boldsymbol{x}) + \log \kappa(\boldsymbol{y})]\, \nu(d\boldsymbol{x}, d\boldsymbol{y})$$

is valid as long as the right hand side is not of the form $\infty - \infty$, which is true by (8.5). The chain rule then gives

$$\begin{aligned}
&- \log \int e^{\frac{1}{2}[\log \kappa(\boldsymbol{x}) + \log \kappa(\boldsymbol{y})]} \bar{\mu}(d\boldsymbol{x}) \varphi(\boldsymbol{x}, d\boldsymbol{y}) \\
&\quad \leq R(\nu \| \bar{\mu} \otimes \varphi) - \int \log \kappa(\boldsymbol{x}) [\nu]_1 (d\boldsymbol{x}) \\
&\quad = R(\nu \| [\nu]_1 \otimes \varphi) + R([\nu]_1 \| \bar{\mu}) - \int \log \kappa(\boldsymbol{x}) [\nu]_1 (d\boldsymbol{x}) \\
&\quad = R(\nu \| [\nu]_1 \otimes \varphi) - \int \log r(\boldsymbol{x}) [\nu]_1 (d\boldsymbol{x}),
\end{aligned}$$

and thus

$$- \int \sqrt{\kappa(\boldsymbol{x}) \kappa(\boldsymbol{y})} \bar{\mu}(d\boldsymbol{x}) \varphi(\boldsymbol{x}, d\boldsymbol{y}) \leq -e^{-R(\nu \| [\nu]_1 \otimes \varphi) + \int \log r(\boldsymbol{x}) [\nu]_1 (d\boldsymbol{x})}.$$

Then (8.4) follows from the fact that if $a \in \mathbb{R}$ and $b \in (0, \infty)$ then $-e^{-a} \leq ab + b \log b - b$ by taking $a = R(\nu \| [\nu]_1 \otimes \varphi) - \int \log r(\boldsymbol{x}) [\nu]_1 (d\boldsymbol{x})$ and $b = A$. □

8.2.3. Decomposition of the exponential distribution. In the construction of η_T^a we used independent exponential random variables $\tau_{i,k}^a$ and geometric random variables N_i^a, and the fact that for each i

$$\sum_{\ell=0}^{N_i^a - 1} \tau_{i,\ell}^a$$

is exponential with mean one. This deomposition corresponds to a relationship in relative entropies, which we now state.

LEMMA 8.3. *For $a \in (0, \infty)$, let $\bar{N}^a$ be distributed according to a random probability measure $\bar{\beta}^a$ on $\{0, 1, \ldots\}$. Given $\bar{N}^a = \ell$, for $k \in \{0, 1, \ldots l\}$ let $\bar{\tau}_k^a(\ell)$ be distributed according to a random probability measure $\bar{\sigma}_k^a(\ell)$ on $[0, \infty)$. Let $\bar{\sigma}$ be the distribution of the random variable*

$$\bar{\tau} = \sum_{k=0}^{\bar{N}^a - 1} \bar{\tau}_k^a(\bar{N}^a).$$

Then

$$E\left[R\left(\bar{\beta}^a \| \beta^a\right) + \sum_{k=0}^{\bar{N}^a - 1} R\left(\bar{\sigma}_k^a(\bar{N}^a) \| \sigma^a\right) \right] \geq E\left[R\left(\bar{\sigma} \| \sigma^1\right) \right].$$

Proof. Define a measure μ on the space $\mathbb{N}_+ \times \prod_{i=0}^{\infty} [0, \infty)$ as follows. For any $\ell \in \mathbb{N}_+$ and any sequence $A_0, A_1, \ldots$ of Borel measurable subsets of $[0, \infty)$,

$$\mu\left(\{l\} \times A_0 \times A_1 \times \ldots\right) = \beta^a(l) \prod_{k=0}^{\infty} \sigma^a(A_k),$$

and similarly define the measure $\bar{\mu}$ by

$$\bar{\mu}\left(\{l\} \times A_0 \times A_1 \times \ldots\right) = \bar{\beta}^a(l) \left(\prod_{k=0}^{\ell - 1} \bar{\sigma}_k^a(\ell)(A_k) \right) \left(\prod_{k=\ell}^{\infty} \sigma^a(A_k) \right).$$

Then by the chain rule of relative entropy

$$E\left[R\left(\bar{\mu}\left\|\mu\right.\right)\right]=E\left[R\left(\bar{\beta}^{a}\left\|\beta^{a}\right.\right)+\sum_{\ell=0}^{\infty}\sum_{k=0}^{\ell-1}R\left(\bar{\sigma}_{k}^{a}(\ell)\left\|\sigma^{a}\right.\right)\bar{\beta}^{a}\left(\ell\right)\right].$$

Since

$$\begin{aligned}E\left[\sum_{k=0}^{\bar{N}^{a}-1}R\left(\bar{\sigma}_{k}^{a}(\bar{N}^{a})\left\|\sigma^{a}\right.\right)\right]&=E\left[E\left[\sum_{k=0}^{\bar{N}^{a}-1}R\left(\bar{\sigma}_{k}^{a}(\bar{N}^{a})\left\|\sigma^{a}\right.\right)\middle|\bar{N}^{a}\right]\right]\\&=E\left[\sum_{\ell=0}^{\infty}\sum_{k=0}^{\ell-1}R\left(\bar{\sigma}_{k}^{a}(\ell)\left\|\sigma^{a}\right.\right)\bar{\beta}^{a}\left(\ell\right)\right],\end{aligned}$$

it follows that

$$E\left[R\left(\bar{\mu}\left\|\mu\right.\right)\right]=E\left[R\left(\bar{\beta}^{a}\left\|\beta^{a}\right.\right)+\sum_{k=0}^{\bar{N}^{a}-1}R\left(\bar{\sigma}_{k}^{a}(\bar{N}^{a})\left\|\sigma^{a}\right.\right)\right].$$

Observe that $\bar{\tau}$ can be written as a measurable mapping on $\mathbb{N}_{+}\times\prod_{i=0}^{\infty}[0,\infty)$ and that $\bar{\sigma}$ and σ^{1} are the distributions induced on $[0,\infty)$ under that map by $\bar{\mu}$ and μ, respectively. Since relative entropy can only decrease under such a mapping, it follows that $ER\left(\bar{\mu}\left\|\mu\right.\right)\geq ER\left(\bar{\sigma}\left\|\sigma^{1}\right.\right)$. □

8.3. Lower bound. The proof of the lower bound will be partitioned into three cases according to a parameter $C\in(1,\infty)$. After the three cases have been argued, the proof of the lower bound will be completed. The first two cases are very simple, and give estimates when R^{a}/T is small (i.e., unusually few exponential clocks with mean 1 are needed before time T is reached) or when R^{a}/T is large (i.e., an unusually large number of of such clocks are needed to reach tome T). The processes $\{\boldsymbol{U}^{a}(j)\}$ and $\{M_{\ell}^{a}(j)\}$ play no role in these estimates, and the required estimates follow from Chebyshev's inequality as it is used in the proof of Crámer's Theorem. We will need the function $h\left(b\right)=-\log b+b-1$, $b\in[0,\infty)$, which satisfies $\inf\left\{R\left(\gamma\left\|\sigma^{1}\right.\right):\int u\gamma\left(du\right)=b\right\}=h\left(b\right)$.

8.3.1. The case $R^{a}/T\leq 1/C$. Let $F:\mathcal{P}(S)\rightarrow\mathbb{R}$ be non-negative and continuous. Then

$$\begin{aligned}&-\frac{1}{T}\log E\left[\exp\{-TF(\eta_{T}^{a})-\infty 1_{\{[0,1/C]\}^{c}}(R^{a}/T)\}\right]\\&\quad=-\frac{1}{T}\log E\left[\exp\left\{-TF(\eta_{T}^{a})\right\}1_{\{[0,1/C]\}}(R^{a}/T)\right]\\&\quad\geq-\frac{1}{T}\log P\left\{\sum_{i=0}^{\lfloor T/C\rfloor+1}\tau_{i}^{1}\geq T\right\}.\end{aligned}$$

By Chebyshev's inequality, for $\alpha\in(0,1)$

$$\begin{aligned}P\left\{\sum_{i=0}^{\lfloor T/C\rfloor+1}\tau_{i}^{1}\geq T\right\}&=P\left\{e^{\alpha\sum_{i=0}^{\lfloor T/C\rfloor+1}\tau_{i}^{1}}\geq e^{\alpha T}\right\}\\&\leq\exp\left\{(\lfloor T/C\rfloor+2)\left(\log\frac{1}{1-\alpha}-\alpha C\right)\right\}.\end{aligned}$$

Optimizing this inequality over $\alpha \in (0,1)$ gives

$$\begin{aligned}\liminf_{T\to\infty} -\frac{1}{T}\log E\left[\exp\{-TF(\eta_T^a) - \infty 1_{\{[0,1/C]\}^c}(R^a/T)\}\right] \\ \geq \liminf_{T\to\infty} \frac{1}{T}\left(\lfloor T/C\rfloor + 2\right) h\left(C\right) \\ = \frac{h\left(C\right)}{C}.\end{aligned}$$

Note that $h\left(C\right)/C \to 1$ as $C \to \infty$.

8.3.2. The case $R^a/T \geq C$.. With F as in the last section, an analogous argument gives

$$\begin{aligned}\liminf_{T\to\infty} -\frac{1}{T}\log E\left[\exp\{-TF(\eta_T^a) - \infty 1_{\{[C,\infty)\}^c}(R^a/T)\}\right] \\ \geq \liminf_{T\to\infty} \frac{1}{T}\left(\lfloor T\left(C-1\right)\rfloor + 2\right) h\left(\frac{1}{C-1}\right) \\ = \left(C-1\right) h\left(\frac{1}{C-1}\right).\end{aligned}$$

Note that $\left(C-1\right) h\left(1/\left(C-1\right)\right) \to \infty$ as $C \to \infty$. .

8.3.3. The case $1/C \leq R^a/T \leq C$.. To analyze this case it will be sufficient to consider any deterministic sequence $\{r^a\}$ such that $1/C \leq r^a/T \leq C$, and such that $r^a/T \to A$ as $T \to \infty$, and obtain lower bounds on

$$\liminf_{T\to\infty} -\frac{1}{T}\log E\left[\exp\{-TF(\eta_T^a) - \infty 1_{\{r^a/T\}^c}(R^a/T)\}\right].$$

Since I^∞ is convex, to prove the lower bound for (8.1) we can assume F is convex and lower semicontinuous [3, Theorem 1.2.1]. The representation from Lemma 8.1 will be applied. We first note that the representation will include a term of the form $\infty 1_{\{r^a/T\}^c}(\bar{R}^a/T)$ on the right hand side. We can remove this term if we restrict the infimum to controls for which $\bar{R}^a = r^a$ w.p.1, and do so for notational convenience. The representation thus becomes

$$\begin{aligned}&-\frac{1}{T}\log E\left[\exp\{-TF(\eta_T^a) - \infty 1_{\{r^a\}^c}(R^a)\}\right] \qquad (8.6)\\ &= \inf E\left[F(\bar{\eta}_T^a) + \frac{1}{T}\sum_{i=0}^{r^a-1}\left[R\left(\bar{\alpha}_i(\bar{\boldsymbol{U}}^a(i),\cdot|\bar{M}_0^a(i))\,\middle\|\,\alpha(\bar{\boldsymbol{U}}^a(i),\cdot|\bar{M}_0^a(i))\right)\right.\right.\\ &\quad \left.+R\left(\bar{\beta}_i^a\,\middle\|\,\beta^a\right)\right] + \frac{1}{T}\sum_{i=0}^{r^a-1}\sum_{k=0}^{\bar{N}_i^a-1}\left[R\left(\bar{p}_{i,k}(\bar{M}_k^a(i),\cdot|\bar{\boldsymbol{U}}^a(i))\,\middle\|\,p(\bar{M}_k^a(i),\cdot|\bar{\boldsymbol{U}}^a(i))\right)\right.\\ &\quad \left.\left.+R\left(\bar{\sigma}_{i,k}^a\,\middle\|\,\sigma^a\right)\right]\right]\end{aligned}$$

There are four relative entropy sums in (8.6). Since F is bounded from below, the lower bound holds vacuously unless each such term is uniformly bounded.

We first show that the empirical distribution on $\bar{N}_i^a$ converges to δ_∞ by using a martingale argument. For $C_1, C_2 \subset [0,\infty)$, let

$$\xi^T(C_1 \times C_2) \doteq \frac{1}{r^a} \sum_{i=0}^{r^a-1} \delta_{\bar{N}_i^a}(C_1)\bar{\beta}_i^a(C_2).$$

Consider the one-point compactification of $[0,\infty)$, which we identify with $[0,\infty]$, and left $f : [0,\infty] \to \mathbb{R}$ be bounded and continuous. Then for any $\varepsilon > 0$

$$\begin{aligned} & P\left\{\left|\int [f(z_1) - f(z_2)]\xi^T(dz_1 \times dz_2)\right| \geq \varepsilon\right\} \qquad (8.7) \\ & = P\left\{\left|\frac{1}{r^a}\sum_{i=0}^{r^a-1}\left(f(\bar{N}_i^a) - \int f(n)\bar{\beta}_i^a(dn)\right)\right| \geq \varepsilon\right\} \\ & \leq \frac{1}{\varepsilon^2} E\left[\left|\frac{1}{r^a}\sum_{i=0}^{r^a-1}\left(f(\bar{N}_i^a) - \int f(n)\bar{\beta}_i^a(dn)\right)\right|^2\right] \\ & \leq \frac{\|f\|_\infty^2}{\varepsilon^2 r^a} \\ & \to 0, \end{aligned}$$

and thus any weak limit of ξ^T has identical marginals. Next note that by Jensen's inequality and since $r^a/T \to A \in (0,\infty)$, the uniform bound on the second relative entropy sum implies that $ER\left([\xi^T]_2 \,\|\beta^a\right)$ is uniformly bounded. Using that $\inf\{R(\gamma\,\|\beta_a) : \int u\gamma(du) = b\} = b\log\frac{b}{a} + (1+b)\log\frac{1+a}{1+b}$, is follows that the weak limit of $[\xi^T]_2 = \delta_\infty$ w.p.1.

Next we consider the asymptotic properties of the collection $\left\{\bar{M}_\ell^a(i)\right\}$, under boundedness of the third relative entropy sum. We use that the empirical measure of the $\left\{\bar{N}_i^a\right\}$ tends to δ_∞. Since $p(\cdot,\cdot|\bar{\boldsymbol{U}}^a(i))$ is the transition function of a finite state Markov chain this means that asymptotically the $\bar{M}_k^a(i), k = 0,\ldots,\bar{N}_i^a$ are samples from the transition probability (8.2) with $\boldsymbol{x} = \bar{\boldsymbol{U}}^a(i)$, and in particular that $\bar{M}_0^a(i+1)$ is asymptotically conditionally independent with distribution $(\rho(\boldsymbol{x}), 1-\rho(\boldsymbol{x}))$. Indeed, a martingale argument similar to (8.7) shows that if

$$\mu^T(C_1 \times C_2) \doteq \frac{1}{r^a}\sum_{i=0}^{r^a-1} \delta_{\bar{\boldsymbol{U}}^a(i)}(C_1)\delta_{\bar{M}_0^a(i)}(C_2),$$

and if μ^∞ is the weak limit of any convergent subsequence (which must exist by compactness), then

$$\left([\mu^\infty]_{2|1}(\{0\}\,|\boldsymbol{y}), [\mu^\infty]_{2|1}(\{1\}\,|\boldsymbol{y})\right) = (\rho(\boldsymbol{y}), 1-\rho(\boldsymbol{y}))\ \ [\mu^\infty]_1\text{-a.s.}, \qquad (8.8)$$

and the same is true for the empirical measure of $\left\{\bar{M}_k^a(i), k = 0,\ldots,\bar{N}_i^a\right\}$.

We next remove the third relative entropy sum from the representation (8.6), and obtain a lower bound for the right hand side using Lemma 8.3. Let $\hat{\sigma}_i^a$ denote the distribution of the random variable

$$\hat{\tau}_i^a = \sum_{\ell=0}^{\bar{N}_i^a-1} \bar{\tau}_{i,\ell}^a$$

Then we have the lower bound

$$-\frac{1}{T}\log E\left[\exp\{-TF(\eta_T^a)-\infty 1_{\{r^a\}^c}(R^a)\}\right] \tag{8.9}$$
$$\geq \inf\left[F(E\bar{\eta}_T^a)+\frac{1}{T}E\sum_{i=0}^{r^a-1}\left[R\left(\bar{\alpha}_i(\bar{\boldsymbol{U}}^a(i),\cdot|\bar{M}_0^a(i))\,\middle\|\,\alpha(\bar{\boldsymbol{U}}^a(i),\cdot|\bar{M}_0^a(i))\right)\right.\right.$$
$$\left.\left.+R\left(\hat{\sigma}_i^a\,\middle\|\,\sigma^1\right)\right]\right].$$

To study the lower bound of (8.9) we introduce the measures

$$\kappa^T(C_1\times C_2\times C_3)\doteq\frac{1}{r^a}\sum_{i=0}^{r^a-1}\delta_{\bar{\boldsymbol{U}}^a(i)}(C_1)\bar{\alpha}_i(\bar{\boldsymbol{U}}^a(i),C_2|\bar{M}_0^a(i))\delta_{\bar{M}_0^a(i)}(C_3)$$
$$\xi^T(C_1\times C_2)\doteq\frac{1}{T}\sum_{i=0}^{r^a-1}\delta_{\bar{\boldsymbol{U}}^a(i)}(C_1)\delta_{(\hat{m}_i^a)^{-1}}(C_2)\hat{m}_i^a,$$

where $\hat{m}_i^a$ is the conditional mean of $\hat{\tau}_i^a$. The restriction on the control measures implies

$$\sum_{i=0}^{r^a-2}\hat{\tau}_i^a\leq T<\sum_{i=0}^{r^a-1}\hat{\tau}_i^a,$$

and since function h is increasing on $(1,\infty)$, we can assume without loss of generality that $\hat{m}_{r^a-1}^a\leq 1$.

We introduce the function $\ell:\mathbb{R}_+\to\mathbb{R}_+$ given by $\ell(b)=b\log b-b+1$. Note that $h(b)=b\ell(1/b)$. The sequence $\{\kappa^T\}$ is tight because S and $\{0,1\}$ are compact. Using the fact that $\hat{\sigma}_i^a$ selects the conditional distribution of $\hat{\tau}_i^a$ and $\inf\{R\left(\gamma\,\middle\|\,\sigma^1\right):\int u\gamma(du)=b\}=h(b)$, we have

$$E\left[\frac{1}{T}\sum_{i=0}^{r^a-1}R\left(\hat{\sigma}_i^a\,\middle\|\,\sigma^1\right)\right]\geq E\left[\frac{1}{T}\sum_{i=0}^{r^a-1}h(\hat{m}_i^a)\right]$$
$$=E\left[\frac{1}{T}\sum_{i=0}^{r^a-1}\hat{m}_i^a\ell([\hat{m}_i^a]^{-1})\right]$$
$$=E\left[\int\ell(z)\xi^T(du\times dz)\right].$$

Hence the uniform bound on the relative entropy sum gives that $\{\xi^T\}$ is tight. Let $f:S\to\mathbb{R}$ be bounded and continuous. Since $\bar{\alpha}_i(\bar{\boldsymbol{U}}^a(i),\cdot|\bar{M}_0^a(i))$ selects the conditional

distribution of $\bar{\boldsymbol{U}}^a(i+1)$, for $\varepsilon > 0$

$$
\begin{aligned}
&P\left\{\left|\sum_{z=1}^{2}\int [f(\boldsymbol{y}_1)-f(\boldsymbol{y}_2)]\kappa^T(d\boldsymbol{y}_1\times d\boldsymbol{y}_2\times dz)\right|\geq\varepsilon\right\}\\
&=P\left\{\left|\frac{1}{r^a}\sum_{i=0}^{r^a-1}\left(f(\bar{\boldsymbol{U}}^a(i+1))-\int f(\boldsymbol{y})\bar{\alpha}_i(\bar{\boldsymbol{U}}^a(i),d\boldsymbol{y}|\bar{M}_0^a(i))\right)\right|\geq\varepsilon\right\}\\
&\leq\frac{1}{\varepsilon^2}E\left[\left|\frac{1}{r^a}\sum_{i=0}^{r^a-1}\left(f(\bar{\boldsymbol{U}}^a(i+1))-\int f(\boldsymbol{y})\bar{\alpha}_i(\bar{\boldsymbol{U}}^a(i),d\boldsymbol{y}|\bar{M}_0^a(i))\right)\right|^2\right]\\
&\leq\frac{\|f\|_\infty^2}{\varepsilon^2 r^a}\\
&\to 0.
\end{aligned}
$$

We find that

$$
[\kappa^\infty]_1=[\kappa^\infty]_2 \text{ w.p.1.} \tag{8.10}
$$

Using Fatou's lemma and the definition of φ, we get the lower bound $AR([\kappa^\infty]_{1,2}\|\,[\kappa^\infty]_1\otimes\varphi)$ on the weak limit of the corresponding relative entropies (the second sum in (8.9)).

With regard to ξ^∞, comparing the form of $\left[\xi^T\right]_1$ and $\bar{\psi}_T^a$ gives

$$
E\left[\xi^\infty\right]_1=E\bar{\psi}_\infty.
$$

Observe that

$$
\int z\xi^T(d\boldsymbol{x}\times dz)=\frac{r^a}{T}\left[\kappa^T\right]_1(d\boldsymbol{x}).
$$

Because of the superlinearity of ℓ we have uniform integrability, and thus passing to the limit gives

$$
\int z[\xi^\infty]_1(d\boldsymbol{x})[\xi^\infty]_{2|1}(dz|\boldsymbol{x})=A[\kappa^\infty]_1(d\boldsymbol{x}).
$$

Hence with the definition $b(\boldsymbol{x})=\int z[\xi^\infty]_{2|1}(dz|\boldsymbol{x})$, $[d[\kappa^\infty]_1/d[\xi^\infty]_1](\boldsymbol{x})=b(\boldsymbol{x})/A$. Using Fatou's lemma, Jensen's inequality and the weak convergence, we get the following lower bound on the corresponding relative entropies (the first sum in (8.9)):

$$
\begin{aligned}
\liminf_{T\to\infty}&\int\ell(z)\xi^T(du\times dz)\\
&\geq\int\ell(z)\xi^\infty(du\times dz)\\
&\geq\int\ell\left(\int z[\xi^\infty]_{2|1}(dz|\boldsymbol{x})\right)[\xi^\infty]_1(d\boldsymbol{x})\\
&=\int\ell\left(b(\boldsymbol{x})\right)[\xi^\infty]_1(d\boldsymbol{x})\\
&=A\int h(1/b(\boldsymbol{x}))[\kappa^\infty]_1(d\boldsymbol{x}).
\end{aligned}
$$

Let $r(\boldsymbol{y}) = A/b(\boldsymbol{y})$. Then we can write the combined lower bound on the relative entropies as

$$\begin{aligned}
&AER([\kappa^\infty]_{1,2}\|[\kappa^\infty]_1 \otimes \varphi) + AE\int h(1/b(\boldsymbol{x}))[\kappa^\infty]_1(d\boldsymbol{x}) \\
&= AER([\kappa^\infty]_{1,2}\|[\kappa^\infty]_1 \otimes \varphi) - AE\int \log r(\boldsymbol{y})[\kappa^\infty]_1(d\boldsymbol{y}) + A\log A - A + 1. \quad (8.11)
\end{aligned}$$

We next consider $\bar{\psi}^a_T$ and $\bar{\eta}^a_T$. Using the limiting properties of the $\bar{M}^a_k(i)$, we have

$$\begin{aligned}
\bar{\eta}^a_T(C) &= \frac{1}{T}\int_0^T \left[1_{\left\{\bar{Z}^a(t)=0\right\}}\delta_{\bar{\boldsymbol{Y}}^a(t)}(C) + 1_{\left\{\bar{Z}^a(t)=1\right\}}\delta_{\bar{\boldsymbol{Y}}^a(t)^R}(C)\right]dt \\
&\to \int \left[\rho(\boldsymbol{y})\delta_{\boldsymbol{y}}(C) + (1-\rho(\boldsymbol{y}))\delta_{\boldsymbol{y}^R}(C)\right][\xi^\infty]_1 \\
&= \bar{\eta}_\infty(C)
\end{aligned}$$

and

$$\bar{\psi}^a_T(C) = \frac{1}{T}\int_0^T \delta_{\bar{\boldsymbol{Y}}^a(t)}(C)dt \to \int_C [\xi^\infty]_1 = \bar{\psi}_\infty(C).$$

Note that this implies the relation

$$\bar{\eta}_\infty(C) = \int_C \left[\rho(\boldsymbol{y})\bar{\psi}_\infty(d\boldsymbol{y}) + \rho(\boldsymbol{y}^R)\bar{\psi}_\infty(d\boldsymbol{y}^R)\right]. \quad (8.12)$$

Finally we consider the weighted empirical measure. By (8.11), Lemma 8.2 and Jensen's inequality we have the lower bound $\left[F(E\bar{\eta}_\infty) + K(E\bar{\psi}_\infty)\right]$ for the limit inferior of the right hand side of (8.6), where $\bar{\eta}_\infty$ and $\bar{\psi}_\infty$ are related by (8.12). Thus we need only show that

$$K(E\bar{\psi}_\infty) \geq I^0(E\bar{\eta}_\infty) = I^\infty(E\bar{\eta}_\infty).$$

The equality follows from the definition of I^∞ and (8.12). Let $a(\boldsymbol{y}) = [dE\bar{\psi}_\infty/d\bar{\mu}](\boldsymbol{y})$. Then

$$K(E\bar{\psi}_\infty) = 1 - \int \sqrt{a(\boldsymbol{x})a(\boldsymbol{y})}\bar{\mu}(d\boldsymbol{x})\varphi(\boldsymbol{x}, d\boldsymbol{y})$$

and

$$I^0(E\bar{\eta}_\infty) = 1 - \int \frac{1}{2}\sqrt{(a(\boldsymbol{x}) + a(\boldsymbol{x}^R))\,(a(\boldsymbol{y}) + a(\boldsymbol{y}^R))}\mu(d\boldsymbol{x})\alpha(\boldsymbol{x}, d\boldsymbol{y}).$$

Using that $\rho(\boldsymbol{x}) = 1 - \rho(\boldsymbol{x}^R)$ and symmetry

$$\begin{aligned}
&\int \sqrt{a(\boldsymbol{x})a(\boldsymbol{y})}\frac{1}{2}\left[\mu(d\boldsymbol{x}) + \mu(d\boldsymbol{x}^R)\right]\left(\rho(\boldsymbol{x})\alpha(\boldsymbol{x}, d\boldsymbol{y}) + \rho(\boldsymbol{x}^R)\alpha(\boldsymbol{x}^R, d\boldsymbol{y}^R)\right) \\
&= \frac{1}{2}\int \left(\sqrt{a(\boldsymbol{x})a(\boldsymbol{y})} + \sqrt{a(\boldsymbol{x}^R)a(\boldsymbol{y}^R)}\right)\mu(d\boldsymbol{x})\left(\rho(\boldsymbol{x})\alpha(\boldsymbol{x}, d\boldsymbol{y}) + \rho(\boldsymbol{x}^R)\alpha(\boldsymbol{x}^R, d\boldsymbol{y}^R)\right) \\
&= \frac{1}{2}\int \left(\sqrt{a(\boldsymbol{x})a(\boldsymbol{y})} + \sqrt{a(\boldsymbol{x}^R)a(\boldsymbol{y}^R)}\right)\mu(d\boldsymbol{x})\alpha(\boldsymbol{x}, d\boldsymbol{y}).
\end{aligned}$$

We now use

$$\sqrt{a(\boldsymbol{x})a(\boldsymbol{y})}+\sqrt{a(\boldsymbol{x}^R)a(\boldsymbol{y}^R)}\leq\sqrt{\left(a(\boldsymbol{x})+a(\boldsymbol{x}^R)\right)\left(a(\boldsymbol{y})+a(\boldsymbol{y}^R)\right)}$$

to obtain $K(E\bar{\psi}_\infty)\geq I^0(E\bar{\eta}_\infty)$. [We note for later use that given $\bar{\eta}$ that is absolutely continuous with respect to Lebesgue measure and such that $I^\infty(\bar{\eta})<\infty$, $K(\bar{\psi})=I^0(\bar{\eta})$ can be shown for a $\bar{\psi}$ that maps to $\bar{\eta}$ by taking $\bar{\psi}=[\bar{\eta}+\bar{\eta}^R]/2$.]

8.3.4. Combining the cases. In the last subsection we showed that for any sequence r^a such that $r^a/T\to A\in[1/C,C]$,

$$\begin{aligned}\liminf_{T\to\infty}-\frac{1}{T}\log E\left[\exp\{-TF(\eta_T^a)-\infty 1_{\{r^a/T\}^c}(R^a/T)\}\right]&\geq\left[F(E\bar{\eta}_\infty)+I^\infty(E\bar{\eta}_\infty)\right]\\&\geq\inf_{\nu\in\mathcal{P}(S)}\left[F(\nu)+I^\infty(\nu)\right],\end{aligned}$$

and an argument by contradiction shows that the bound is uniform in A. Thus

$$\begin{aligned}&\liminf_{T\to\infty}-\frac{1}{T}\log\left\{\sum_{r^a=\lceil\frac{T}{C}\rceil}^{\lfloor TC\rfloor}E\left[\exp\{-TF(\eta_T^a)-\infty 1_{\{r^a/T\}^c}(R^a/T)\}\right]\right\}\\&\quad\geq\liminf_{T\to\infty}-\frac{1}{T}\log\left\{TC\cdot\bigvee_{r^a=\lceil\frac{T}{C}\rceil}^{\lfloor TC\rfloor}E\left[\exp\{-TF(\eta_T^a)-\infty 1_{\{r^a/T\}^c}(R^a/T)\}\right]\right\}\\&\quad\geq\inf_{\nu\in\mathcal{P}(S)}\left[F(\nu)+I^\infty(\nu)\right].\end{aligned}$$

We now partition $E\left[\exp\{-TF(\eta_T^a)\right]$ according to the various cases to obtain the overall lower bound

$$\begin{aligned}&\liminf_{T\to\infty}-\frac{1}{T}\log E\left[\exp\{-TF(\eta_T^{a_T})\}\right]\\&\qquad\geq\min\left\{\inf_{\nu\in\mathcal{P}(S)}\left[F(\nu)+I^\infty(\nu)\right],\frac{h\left(C\right)}{C},\left(C-1\right)h\left(\frac{1}{C-1}\right)\right\}.\end{aligned}$$

Letting $C\to\infty$ and using the fact that $I^\infty\leq 1$, we have the desired lower bound

$$\liminf_{T\to\infty}-\frac{1}{T}\log E\left[\exp\{-TF(\eta_T^{a_T})\}\right]\geq\inf_{\nu\in\mathcal{P}(S)}\left[F(\nu)+I^\infty(\nu)\right].$$

8.4. Upper bound. The proof of the reverse inequality

$$\limsup_{T\to\infty}-\frac{1}{T}\log E\left[\exp\{-TF(\eta_T^{a_T})\}\right]\leq\inf_{\nu\in\mathcal{P}(S)}\left[F(\nu)+I^\infty(\nu)\right]\tag{8.13}$$

is simpler. Let bounded and continuous F be given. Given $\varepsilon>0$, we can find ν that is absolutely continuous with respect to Lebesgue measure and for which

$$\left[F(\nu)+I^\infty(\nu)\right]\leq\inf_{\nu\in\mathcal{P}(S)}\left[F(\nu)+I^\infty(\nu)\right]+\varepsilon.$$

Next we use that I^∞ is convex and $I^\infty(\mu)=0$ to find $\tau>0$ such that

$$\left[F(\nu^\tau)+I^\infty(\nu^\tau)\right]\leq\left[F(\nu)+I^\infty(\nu)\right]+\varepsilon$$

when $\nu^\tau = \tau\mu + (1-\tau)\nu$. Note that if $\theta(\boldsymbol{x}) = [d\nu/d\mu](\boldsymbol{x})$ and $\theta^\tau(\boldsymbol{x}) = [d\nu^\tau/d\mu](\boldsymbol{x})$, then $\theta^\tau(\boldsymbol{x}) \geq \tau > 0$ and so $I^\infty(\nu^\tau) < 1$.

We will construct a control to use in the representation such $\bar{\eta}_T^{a_T}$ will converge w.p.1 to ν^τ and RE^T will converge w.p.1 to $I^\infty(\nu^\tau)$. These convergences will follow from the ergodic theorem, and the bound $\theta^\tau(\boldsymbol{x}) \geq \tau$ will be used to argue that the ergodicity of the original process is inherited by the controlled process.

We now proceed to the construction. Let $\varsigma(d\boldsymbol{x}) = [\nu^\tau(d\boldsymbol{x}) + \nu^\tau(d\boldsymbol{x}^R)]/2$, so that $I^\infty(\nu^\tau) = K(\varsigma)$. Let $\kappa(\boldsymbol{x}) = [d\varsigma/d\bar{\mu}](\boldsymbol{x}) \geq \tau/2$. Then

$$K(\varsigma) = 1 - \int \sqrt{\kappa(\boldsymbol{x})\kappa(\boldsymbol{y})}\bar{\mu}(d\boldsymbol{x})\varphi(\boldsymbol{x}, d\boldsymbol{y}).$$

Since κ is uniformly bounded from below we have the dual relationship

$$\begin{aligned}
&-\log \int \sqrt{\kappa(\boldsymbol{x})\kappa(\boldsymbol{y})}\bar{\mu}(d\boldsymbol{x})\varphi(\boldsymbol{x}, d\boldsymbol{y}) \\
&= -\log \int e^{\frac{1}{2}[\log\kappa(\boldsymbol{x}) + \log\kappa(\boldsymbol{y})]}\bar{\mu}(d\boldsymbol{x})\varphi(\boldsymbol{x}, d\boldsymbol{y}) \\
&= R(\eta \| \bar{\mu} \otimes \varphi) - \frac{1}{2}\int [\log\kappa(\boldsymbol{x}) + \log\kappa(\boldsymbol{y})]\, \nu(d\boldsymbol{x}, d\boldsymbol{y}),
\end{aligned}$$

where

$$\eta(C) = \frac{\int_C e^{\frac{1}{2}[\log\kappa(\boldsymbol{x}) + \log\kappa(\boldsymbol{y})]}\bar{\mu}(d\boldsymbol{x})\varphi(\boldsymbol{x}, d\boldsymbol{y})}{\int_{S^2} e^{\frac{1}{2}[\log\kappa(\boldsymbol{x}) + \log\kappa(\boldsymbol{y})]}\bar{\mu}(d\boldsymbol{x})\varphi(\boldsymbol{x}, d\boldsymbol{y})}.$$

Since $\sqrt{\kappa(\boldsymbol{x})\kappa(\boldsymbol{y})}$ is symmetric in $\boldsymbol{x}$ and $\boldsymbol{y}$ automatically $[\eta]_1 = [\eta]_2$, and using the bound $\kappa(\boldsymbol{x}) \geq \tau/2$ we can factor $\eta(d\boldsymbol{x} \times d\boldsymbol{y}) = [\eta]_1(d\boldsymbol{x})\bar{\varphi}(\boldsymbol{x}, d\boldsymbol{y})$, where $\bar{\varphi}$ is the transition kernel of a uniformly ergodic Markov chain. Let

$$\bar{\alpha}(\boldsymbol{x}, d\boldsymbol{y}|0) = \bar{\alpha}(\boldsymbol{x}, d\boldsymbol{y}|1) = \bar{\varphi}(\boldsymbol{x}, d\boldsymbol{y}).$$

Notice that from the definition of $\bar{\varphi}$, $\bar{\varphi}(\boldsymbol{x}, d\boldsymbol{y}) = \bar{\varphi}(\boldsymbol{x}^R, d\boldsymbol{y}^R)$, hence $\bar{\alpha}$ has the property that

$$\bar{\alpha}(\boldsymbol{x}^R, d\boldsymbol{y}^R|0) = \bar{\alpha}(\boldsymbol{x}, d\boldsymbol{y}|1)$$

These will be the transition kernels used to construct the $\bar{\boldsymbol{U}}^a(i)$.

Now of course the invariant distribution of $\bar{\varphi}$ is $[\eta]_1$, and not the desired distribution ς. Let $r(\boldsymbol{x}) = [d\varsigma/d[\eta]_1](\boldsymbol{x})$. Then $r(\boldsymbol{x})$ identifies the way in which the distribution of the random variables $\bar{N}_j^a$ should be modified so that in the continuous time the empirical measure $[\eta]_1$ is reshaped into ς. Choose A so that equality holds in (8.4), and set $b(\boldsymbol{x}) = Ar(\boldsymbol{x})$. Then $\bar{\beta}_i^a$ is chosen to be $\beta^{ab(\boldsymbol{x})}$. Consistent with the analysis of the upper bound, we do not perturb the distribution of the other variables, so that $\bar{p}_{i,k}(\cdot, \cdot|\cdot) = p(\cdot, \cdot|\cdot)$ and $\bar{\sigma}_{j,\ell}^a = \sigma^a$. We then construct the controlled processes using these measures in exact analogy with the construction of the original process.

With this choice and Lemma 8.1 we obtain the bound

$$
\begin{aligned}
&-\frac{1}{T}\log E\left[\exp\{-TF(\eta_T^{a_T})\}\right] \\
&\le E\left[F(\bar\eta_T^a)+\frac{1}{T}\sum_{i=0}^{\bar R^a-1}\left[R\left(\bar\alpha(\bar{\boldsymbol U}^a(i),\cdot|\bar M_0^a(i))\,\middle\|\,\alpha(\bar{\boldsymbol U}^a(i),\cdot|\bar M_0^a(i))\right)\right.\right. \\
&\left.\left.+R\left(\beta^{ab(\bar{\boldsymbol U}^a(i))}\,\middle\|\,\beta^a\right)\right]\right].
\end{aligned}
$$

Now apply the ergodic theorem, and use that w.p.1 $A^T=\bar R^a/T\to 1/\int b(\boldsymbol x)[\eta]_1(d\boldsymbol x)$ $=A$, and also that asymptotically the conditional distribution of $\bar M_0^a(i)$ is given by $(\rho_0(\bar{\boldsymbol U}^a(i)),\rho_1(\bar{\boldsymbol U}^a(i)))$. Then right hand side of the last display converges to we have the limit

$$
\begin{aligned}
&F(\nu^\tau)+\frac{1}{\int b(\boldsymbol x)[\eta]_1(d\boldsymbol x)}\int\left[\sum_{z=1}^{2}R\left(\bar\alpha(\boldsymbol x,d\boldsymbol y|z)\,\|\alpha(\boldsymbol x,d\boldsymbol y|z)\right)\rho_z(\boldsymbol x)\right]r(\boldsymbol x)[\eta]_1(d\boldsymbol x) \\
&\quad+\frac{1}{\int b(\boldsymbol x)[\eta]_1(d\boldsymbol x)}\int\left(-\log b(\boldsymbol x)+b(\boldsymbol x)-1\right)b(\boldsymbol x)[\eta]_1(d\boldsymbol x) \\
&=F(\nu^\tau)+A\int R\left(\bar\varphi(\boldsymbol x,d\boldsymbol y)\,\|\varphi(\boldsymbol x,d\boldsymbol y)\right)\varsigma(d\boldsymbol x)-A\int\log r(\boldsymbol x)\varsigma(d\boldsymbol x)+A\log A-A+1 \\
&=F(\nu^\tau)+K(\varsigma) \\
&=F(\nu^\tau)+I^\infty(\nu^\tau).
\end{aligned}
$$

This completes the proof of (8.13) and also the proof of Theorem 4.1.

REFERENCES

[1] L. Breiman. *Probability Theory.* Addison-Wesley, Reading, Mass., 1968.

[2] M.D. Donsker and S.R.S. Varadhan. Asymptotic evaluation of certain Markov process expectations for large time, I. *Comm. Pure Appl. Math.*, 28:1–47, 1975.

[3] P. Dupuis and R. S. Ellis. *A Weak Convergence Approach to the Theory of Large Deviations.* John Wiley & Sons, New York, 1997.

[4] D.J. Earl and M.W. Deem. Parallel tempering: theory, applications, and new perspectives. *Phys. Chem. Chem. Phys.*, 7:3910–3916, 2005.

[5] S.N. Ethier and T.G. Kurtz. *Markov Processes: Characterization and Convergence.* Wiley, New York, 1986.

[6] C.J. Geyer. Markov chain Monte Carlo maximum likelihood. In *Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface*, pages 156–163, New York, 1991. American Statistical Association.

[7] C. Kipnis and S.R.S. Varadhan. Central limit theorem for additive functionals of reversible Markov processes and applications to simple exclusions. *Comm. Math. Phys.*, 104:1–19, 1986.

[8] D.A. Kofke. On the acceptance probability of replica-exchange Monte Carlo trials. *J. Chem. Phys.*, 117:6911–6914, 2002.

[9] H. J. Kushner. *Approximation and Weak Convergence Methods for Random Processes with Applications to Stochastic System Theory.* MIT Press, Cambridge, MA, 1984.

[10] J.S. Liu. *Monte Carlo Strategies in Scientific Computing.* Springer, New York, 2004.

[11] S.P. Meyn and R.L. Tweedie. *Markov Chains and Stochastic Stability.* Springer-Verlag, London, 1993.

[12] P.H. Peskun. Optimum Monte Carlo sampling using Markov chains. *Biometrika*, 60:607–612, 1973.

[13] R. Pinsky. On evaluating the Donsker-Varadhan I-function. *Ann. Probab.*, 13:342–362, 1985.
[14] N. Plattner, J.D. Doll, P. Dupuis, H. Wang, Y. Liu, and J.E. Gubernatis. An infinite swapping approach to the rare-event sampling problem. *J. of Chemical Physics*, 135:134111, 2011.
[15] C. Predescu, M. Predescu, and C.V. Ciobanu. On the efficiency of exchange in parallel tempering Monte Carlo simulations. *J. Phys. Chem. B*, 109:4189–4196, 2005.
[16] C. Presescu, M. Presescu, and C.V. Ciobanu. The incomplete beta function law for parallel tempering sampling of classical canonical systems. *J. Phys. Chem.*, 120:4119–4128, 2004.
[17] J. Rosenthal. Asymptotic variance and convergence rates of nearly periodic MCMC algorithms. *J. Amer. Stat. Asso.*, 98:169–177, 2003.
[18] P.J. Rossky, J.D. Doll, and H.L. Friedman. Brownian dynamics as smart Monte Carlo. *J. Chem. Phys.*, 69:4628–4633, 1978.
[19] D. Sindhikara, D.J. Emerson, and A.E. Roitberg. Exchange often and properly in replica exchange molecular dynamics. *Journal of Chemical Theory and Computation*, 6:2804–2808, 2010.
[20] D. Sindhikara, Y. Meng, and A.E. Roitberg. Exchange frequency in replica exchange molecular dynamics. *The Journal of Chemical Physics*, 128:024103, 2008.
[21] Y. Sugita and Y. Okamoto. The incomplete beta function law for parallel tempering sampling of classical canonical systems. *Chem. Phys. Lett.*, 314:141–151, 1999.
[22] R.H. Swendsen and J.S. Wang. Replica Monte Carlo simulation of spin glasses. *Phys. Rev. Lett.*, 57:2607–2609, 1986.
[23] L. Tierney. A note on Metropolis-Hastings kernels for general state space. *Ann. Appl. Prob.*, 8:1–9, 1998.
[24] D. Wales. *Energy Landscapes: Applications to Clusters, Biomolecules and Glasses*. Cambridge University Press, UK, 2003.
[25] D.B. Woodard, S.C. Schmidler, and M. Huber. Conditions for rapid mixing of parallel and simulated tempering on multimodal distributions. *Ann. Appl. Prob.*, 19:617–640, 2009.